UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

CORSO DI LAUREA MAGISTRALE IN INFORMATICA

DATA SCIENCE E MACHINE LEARNING



PROGETTO DI STATISTICA E ANALISI DEI DATI

**Violenza di genere durante la pandemia: aggravamento o attenuazione dei numeri?**

|  |  |
| --- | --- |
| DOCENTE  Prof. Amelia Giuseppina Nobile | STUDENTI  Maria Natale, matricola: 0522500967  Gaetano Casillo, matricola:????????? |

ANNO ACCADEMICO 2020-2021

Sommario

[1 Introduzione 4](#_Toc58596987)

[1.1 Caso di studio 4](#_Toc58596988)

[1.2 Rappresentazione grafica 6](#_Toc58596989)

[2 Statistica descrittiva univariata 8](#_Toc58596990)

[2.1 Funzione di distribuzione empirica continua 8](#_Toc58596991)

[2.1.1 Funzione di distribuzione empirica continua Utenti 8](#_Toc58596992)

[2.1.2 Funzione di distribuzione empirica continua Vittime 9](#_Toc58596993)

[2.2 Indici di sintesi 10](#_Toc58596994)

[2.2.1 Indici di sintesi Utenti 11](#_Toc58596995)

[2.2.2 Indici di sintesi Vittime 16](#_Toc58596996)

[2.3 Forma della distribuzione di frequenze 19](#_Toc58596997)

[2.3.1 Forma della distribuzione di frequenze Utenti 21](#_Toc58596998)

[2.3.2 Forma della distribuzione di frequenze Vittime 21](#_Toc58596999)

[3 Statistica descrittiva bivariata 22](#_Toc58597000)

[3.1 Regressione lineare semplice 23](#_Toc58597001)

[3.1.1 Regressione lineare semplice Utenti 24](#_Toc58597002)

[3.1.2 Regressione lineare semplice Vittime 28](#_Toc58597003)

[3.2 Regressione lineare multipla 32](#_Toc58597004)

[3.2.1 Regressione lineare multipla Utenti 32](#_Toc58597005)

[3.2.2 Regressione lineare multipla Vittime 36](#_Toc58597006)

[4 Analisi dei cluster 40](#_Toc58597007)

[4.1 Suddivisione in cluster Utenti 42](#_Toc58597008)

[4.2 Suddivisione in cluster Vittime 48](#_Toc58597009)

[5 Variabile aleatoria esponenziale 54](#_Toc58597010)

[5.1 Stima del parametro non noto λ 57](#_Toc58597011)

[5.1.1 Stima puntuale 57](#_Toc58597012)

[5.1.2 Stima intervallare 58](#_Toc58597013)

[5.1.3 Confronto tra due popolazioni esponenziali 60](#_Toc58597014)

[5.2 Verifica delle ipotesi 62](#_Toc58597015)

[6 Bibliografia 63](#_Toc58597016)

# Introduzione

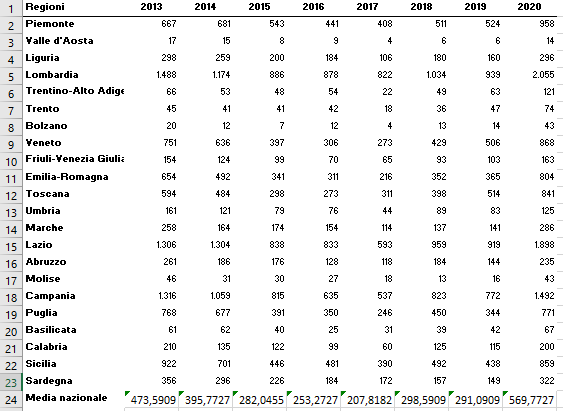
Si sente spesso parlare di violenza di genere, ma che cosa vuol dire? Le Nazioni Unite hanno definito la violenza di genere come “*ogni atto legato alla differenza di sesso che provochi o possa provocare un danno fisico, sessuale, psicologico o una sofferenza della donna, compresa la minaccia di tali atti, la coercizione o l’arbitraria privazione della libertà sia nella vita pubblica che nella vita privata*” [1]. Per supportare le vittime delle violenze di genere è stato attivato il numero verde 1522 attivo 24 ore su 24 offrendo accoglienza in italiano, inglese, francese, spagnolo e arabo.

## Caso di studio

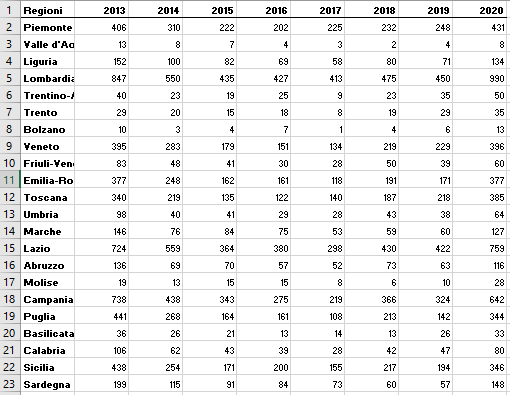
Nel 2020 è stato vissuto il lockdown per 3 mesi, in questo periodo molto si è parlato del lato economico, della scuola, ma poco si è discusso del lato sociale di questo evento. Si è pensato pertanto di analizzare le chiamate e i messaggi effettuate al 1522, numero verde contro lo stalking e la violenza, confrontandole con lo stesso periodo (marzo/giugno) degli anni precedenti. In particolare, nell’analisi statistica effettuata si fa riferimento agli utenti e alle vittime per regione di provenienza e anno. Secondo quanto rilevato dall’Istat, che ha analizzato i dati messi a disposizione dal numero antiviolenza 1522, tra marzo e giugno 2020 le telefonate e le comunicazioni via chat con il centralino sono [più che raddoppiate](https://www.istat.it/it/archivio/246557) rispetto allo stesso periodo dell’anno precedente, con un +119,6%. Il numero complessivo di contatti validi nel periodo preso in considerazione è stato di 15.280, di cui circa un terzo poi trasferiti ad altri servizi come quello dei centri antiviolenza. Numeri assolutamente non paragonabili col passato [2].

Per l’analisi del fenomeno in esame si considerano i dati relativi agli utenti del numero antiviolenza 1522 effettuate nei mesi di marzo-giugno suddivisi per regione ed anno (2013-2020). In particolare, nell’analisi statistica univariata, verranno esaminate nei dettagli le curve relativi ai dati della regione Campania e la media delle chiamate degli utenti e delle vittime effettuate sull’intero territorio nazionale.

Nella seguente tabella vengono mostrati i dati relativi agli utenti del numero 1522 suddivisi per regione ed anno.



Nella seguente tabella vengono mostrati i dati relativi alle vittime del numero 1522 suddivisi per regione ed anno.



## Rappresentazione grafica

Di seguito vengono mostrati i due barplot relativi ai dati della Campania e della media sull’intero territorio nazionale per quanto riguarda la tabella Utenti. In entrambi i casi si può notare che la modalità a cui è associata la frequenza più alta è il 2020.

Il codice per realizzare il barplot della Campania:

png("grafici/chiamateEffettuateUtentiCampaniaFrequenza.png")

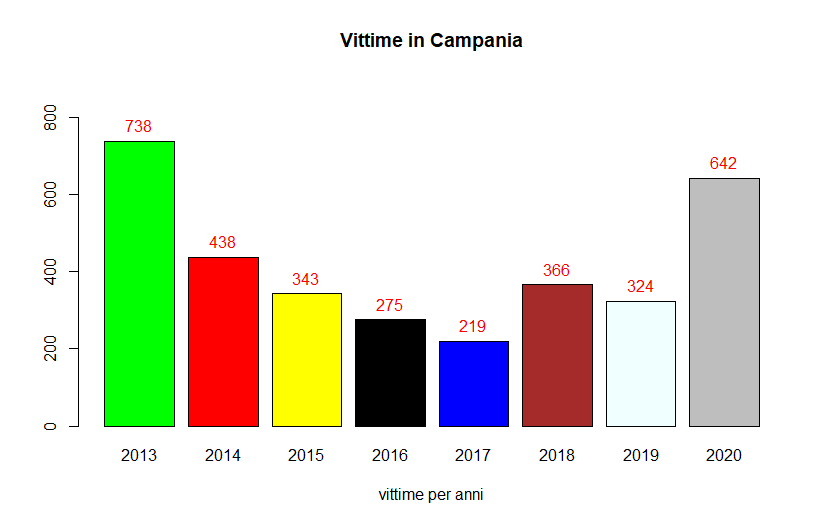
x<-barplot(utenti\_campania, xlab="Anni", ylab="Numero di chiamate effettuate", ylim=c(0,1800), col=1:9,

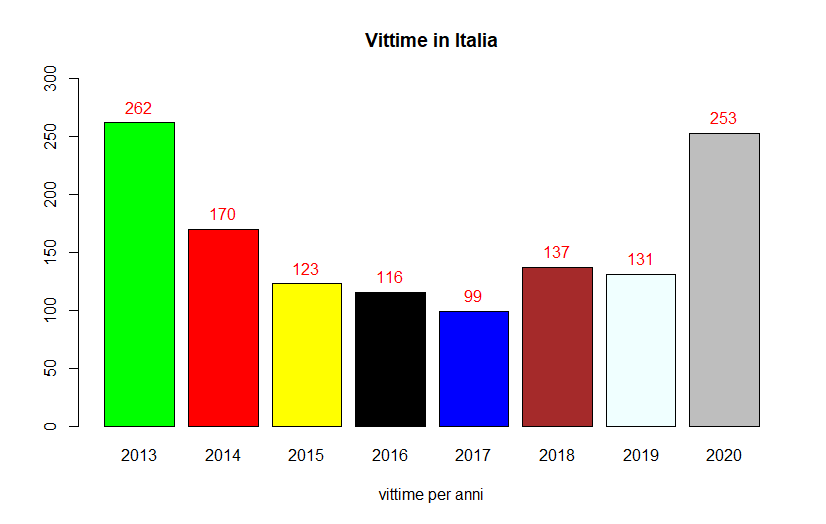
names.arg = colnames, main ="Numero di utenti al numero verde in Campania")

text(x, y=utenti\_campania, pos = 3, labels = utenti\_campania, col="red")

dev.off()

Di seguito vengono mostrati i due barplot relativi ai dati della Campania e della media sull’intero territorio nazionale per quanto riguarda la tabella Vittime. In entrambi i casi si può notare che la modalità a cui è associata la frequenza più alta è il 2013, ma nel 2020, dopo diversi anni di decrescita del numero di vittime, si è avuto un aumento pari al doppio delle vittime del 2019.





# Statistica descrittiva univariata

In questo capitolo verranno mostrati i risultati relativi all’analisi statistica univariata. In particolare, verrà mostrata la funzione di distribuzione empirica continua, i valori degli indici di sintesi, i quartili calcolati con i differenti algoritmi di R e gli indici di dispersione. Infine, verrà analizzata la forma della distribuzione di frequenze attraverso il calcolo della skewness campionaria e della curtosi campionaria. Le varie analisi verranno effettuate prendendo in esame i dati della Campania e della media nazionale negli anni 2013-2020, analizzando le tabelle Utenti e Vittime.

## Funzione di distribuzione empirica continua

La funzione di distribuzione empirica continua viene utilizzata nel caso di dati continui che vengono strutturati in classi. Ad esempio, se si vuole considerare k classi distinte, le classi saranno così caratterizzate: C1 = [z0, z1), C2 = [z1, z2), … Ck=[zk-1, zk] con z0 < z1 < … < zk-1 < zk, dove z0 corrisponde al minimo delle osservazioni e zk corrisponde al massimo delle osservazioni. La funzione di distribuzione empirica continua viene calcolata a partire dalle frequenze relative cumulative associate alle varie classi.

### Funzione di distribuzione empirica continua Utenti

Per calcolare la funzione di distribuzione continua relativa alla tabella Utenti le osservazioni sono state suddivise in tre classi. Per quanto riguarda la media nazionale le classi individuate sono le seguenti: C1 = [208, 329), C2 = [329, 450), C3 = [450, 570]. Per quanto riguarda la Campania le classi individuate sono le seguenti: C1 = [537, 855), C2 = [855, 1173), C3 = [1173, 1492].

Il seguente codice mostra come sono state calcolate le frequenze relative cumulative per le classi individuate per quanto riguarda la Campania.

minOsservazione = min(utenti\_campania)

maxOsservazione = max(utenti\_campania)

frequenza<-table(utenti\_campania)/length(utenti\_campania)

lung<-length(frequenza)

classe<-round((maxOsservazione-minOsservazione)/3, digits=0)

classi<-c(minOsservazione, minOsservazione+classe, minOsservazione+2\*classe, maxOsservazione)

frelclassi <-table (cut (utenti\_campania, breaks = classi,right = FALSE ))/ length (utenti\_campania)

Fcum <-cumsum (frelclassi)

Fcum[3]<-Fcum[3]+frequenza[lung]

Dopo aver calcolato le frequenze relative cumulative. sono stati quindi creati i grafici che mostrano le frequenze di distribuzione continue della Campania e dell’intera nazione.





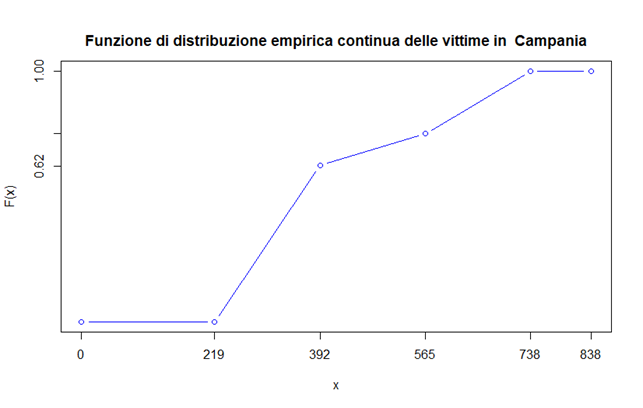
### Funzione di distribuzione empirica continua Vittime

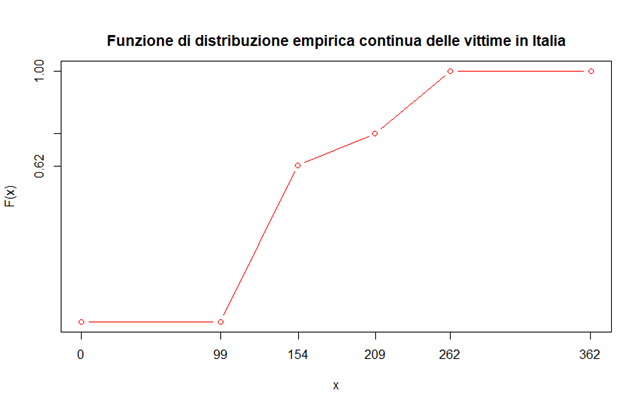
Per calcolare la funzione di distribuzione continua le osservazioni sono state divise in tre classi:

Per la Campania: C1 =[219,392), C2 =[392,565), C3 = [565,738]

Per la media nazionale: C1 =[99,154), C2 = [154,209), C3 = [209,262]

Di seguito verranno mostrati grafici che mostrano le frequenze di distribuzione continua in Campania e in Italia.





## Indici di sintesi

Alcuni indici di sintesi utili a descrivere i dati sono media, mediana, moda, varianza, deviazione standard e coefficiente di variazione. Le prime tre sono misure di centralità dei dati mentre le altre misurano la loro dispersione.

Supponiamo di avere un insieme, x1 , x2 ,…, xn di n valori numerici.

Si definisce **media campionaria** la quantità:

Dato un campione di dati ordinato in maniera crescente, si definisce la **mediana** (o **valore mediano**) come il valore/modalità assunto dalle unità statistiche che si trovano nel mezzo della distribuzione. Se n è dispari, la mediana sarà il valore in posizione (n+1)/2; se n è pari la mediana sarà la media aritmetica dei valori in posizione n/2 e n/2+1. La mediana, quindi è quel valore che divide a metà l’insieme dei dati ordinati. Oltre a questo indice si possono considerare altri indici di posizione detti quantili che consentono di suddividere l’insieme dei dati ordinati in un fissato numeri di parti uguali. In particolare, verranno considerati i quartili che consentono di dividere l’insieme dei dati ordinati in quattro parti uguali.

La **moda campionaria** di un insieme di dati è il valore a cui è associata la frequenza più elevata, non è obbligatorio che la moda esista in ogni insieme di dati e se esiste, è possibile che ne esista più di una; in questo caso, ogni valore è detto “valore modale”.

Avendo un insieme di dati numerici (), si definisce **varianza campionaria** e si indica con , la quantità:

Si definisce **deviazione standard campionaria** la radice quadrata della varianza ossia:

Assegnato un campione di dati numerici x1 , x2 ,…, xn , si definisce **coefficiente di variazione** il rapporto tra la deviazione standard campionaria e il modulo della media campionaria: .

### Indici di sintesi Utenti

Nel grafico seguente vengono mostrate le due curve relative ai dati che si stanno analizzando.



Entrambe le curve mostrano una distribuzione di frequenze non simmetrica, in particolare inizialmente sono decrescenti ed hanno un picco massimo nell’ultimo anno 2020. Le due curve sono tra loro abbastanza simili.

Il grafico seguente mostra, invece, i boxplot di entrambi i campioni di dati per illustrare alcune caratteristiche della distribuzione di frequenza come centralità, dispersione, forma e la presenza di eventuali valori anomali. Il boxplot, detto anche “scatola con i baffi”, rappresenta una scatola i cui estremi sono Q1 (primo quartile)e Q3 (terzo quartile) tagliata da una linea orizzontale in corrispondenza di Q2 (secondo quartile). Sono inoltre presenti due ulteriori linee che rappresentano i baffi in alto e in basso. Il baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale a , mentre il baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a . Se tutti i dati rientrano nell’intervallo , i baffi risultano essere posti in corrispondenza del minimo e del massimo dei dati del campione. I valori anomali al di fuori di tale intervallo vengono visualizzati sotto forma di punti nel grafico.



Entrambi i boxplot rivelano la presenza di asimmetria nei dati in quanto le distanze tra primo e terzo quartile dalla linea della mediana sono molto diverse tra loro. Si può intuire che le curve hanno una coda più allungata a destra e ciò verrà confermato attraverso il calcolo della skewness campionaria. Inoltre, non viene rilevata la presenza di valori anomali.

Utilizzando la funzione summary in R è possibile calcolare minimo, massimo, media, mediana, primo e terzo quartile. In generale, la curva dei due campioni di dati è sostanzialmente simile anche se i dati relativi all’intera nazione sono più bassi in quanto sono ottenuti dalla media di tutte le nazioni, che viene fortemente influenzata dai valori bassi presenti in molte regioni con meno abitanti rispetto alla Campania.





Per individuare la moda si considerano gli istogrammi delle frequenze dei dati considerando la loro suddivisione nelle seguenti classi: C1 = [0, 500), C2 = [500, 1000), C3 = [1000, 1500) C4= [1500, 2000), C5= [2000, 2500].

La classe modale per l’Italia è C1 = [0, 500), in particolare tutti i valori sono concentrati in quella classe. Per la Campania invece la classe modale risulta essere C2 = [500, 1000).



Codice per realizzare gli istogrammi delle frequenze delle classi:

classi<-c(0, 500, 1000, 1500, 2000, 2500)

fclassiCampania <-table (cut (utenti\_campania, breaks = classi,right = FALSE, dig.lab = 10))

for (i in 1:length(utenti\_campania)){

if(utenti\_campania[i]==2500)

fclassiCampania[3]<-fclassiCampania[3]+1

}

fclassiItalia <-table (cut (utenti\_nazione, breaks = classi,right = FALSE, dig.lab=10))

for (i in 1:length(utenti\_nazione)){

if(utenti\_nazione[i]==2500)

fclassiItalia[3]<-fclassiItalia[3]+1

}

png("grafici/istogrammaClassiCampania.png")

hist(utenti\_campania, breaks=classi, col=rainbow(3), main="Istogramma delle frequenze delle classi in Campania")

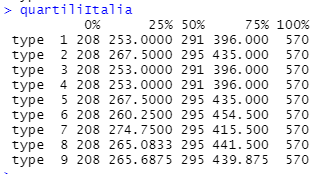
dev.off()

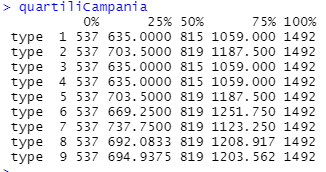
png("grafici/istogrammaClassiItalia.png")

hist(utenti\_nazione, breaks=classi, col=rainbow(3), main="Istogramma delle frequenze delle classi in Italia")

dev.off()

In R ci sono 9 diversi tipi di algoritmi che consentono di calcolare i quartili. Di seguito vengono mostrati i risultati ottenuti dal calcolo dei quartili per quanto riguarda gli utenti della Campania e della media nazionale.





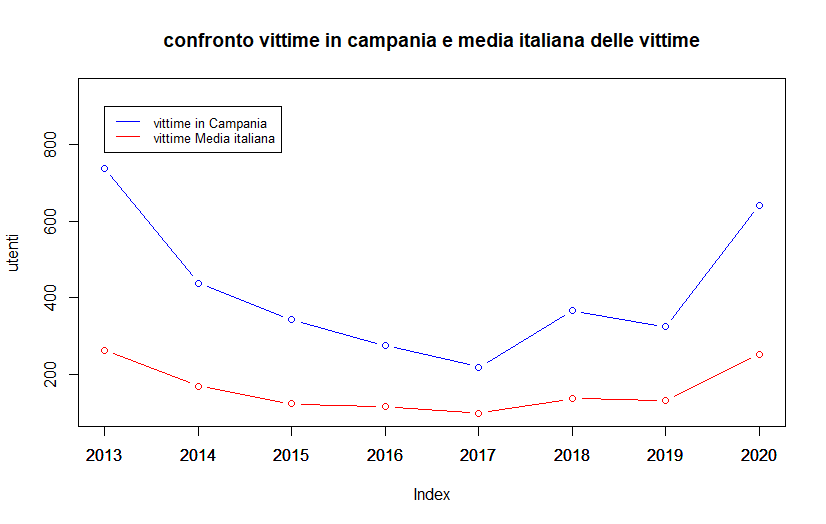
Il seguente codice permette di mostrare invece varianza, deviazione standard e coefficiente di variazione dei due campioni di dati.



Il coefficiente di variazione del campione di dati della Campania è circa 0.3567, mentre quello della media nazionale è circa 0.3551. I due coefficienti sono tra loro molto vicini, indicano quindi una dispersione dei dati attorno alla media molto simile.

### Indici di sintesi Vittime

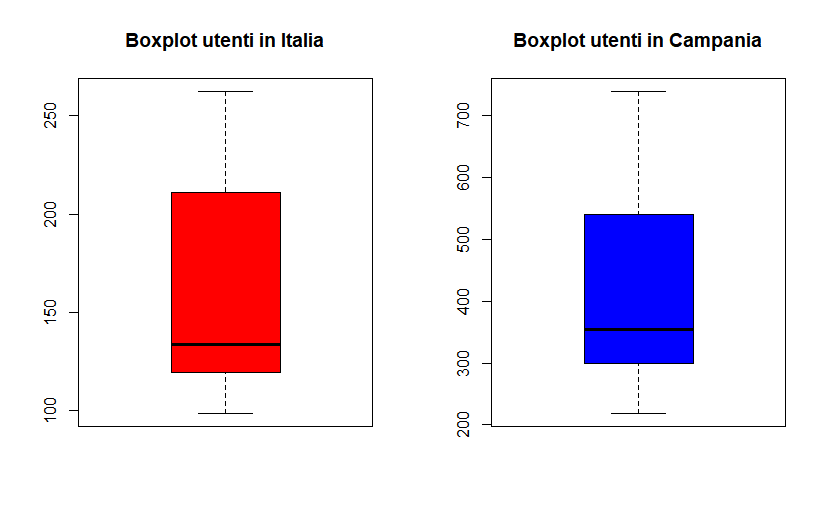
Di seguito è riportato il grafico che rappresenta le curve dei dati che si stanno analizzando, è stata preferenza del programmatore rappresentare le due curve in un solo grafo per mostrare meglio la differenza di numeri, ma andamento simile tra le vittime in media nazionale e le vittime in Campania



Il picco è presente in entrambi i casi nel 2013, per poi avere un andamento discendente fino al 2017 (anno del me too), per poi risalire dal 2018 e arrivare ad un incremento vertiginoso nel 2020.

È da ricordare che i numeri che si stanno analizzando fanno parte di denunce da donne, quindi, è lecito pensare che un grosso movimento quale il me too abbia dato coraggio alle donne che ricevevano abusi di farsi avanti e denunciare i propri aguzzini, nel 2020 l’obbligo della convivenza forzata ha solo incrementato quello che era già presente negli anni precedenti.

Di seguito verranno mostrati alcuni dettagli tramite boxplot riguardo le distribuzioni di frequenza e i vari quartili, la mediana e la media.



> summary(vittimecampania)

"Min. 1st Qu. Median Mean 3rd Qu. Max.

219.0 311.8 354.5 418.1 489.0 738.0

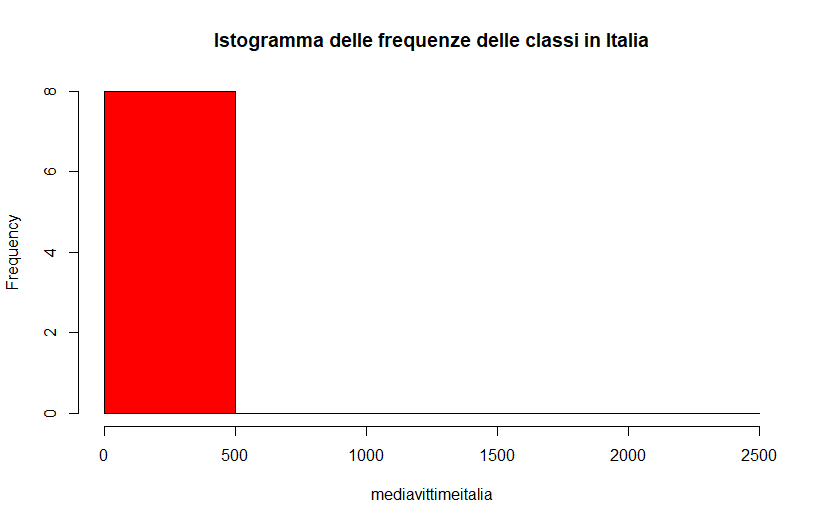
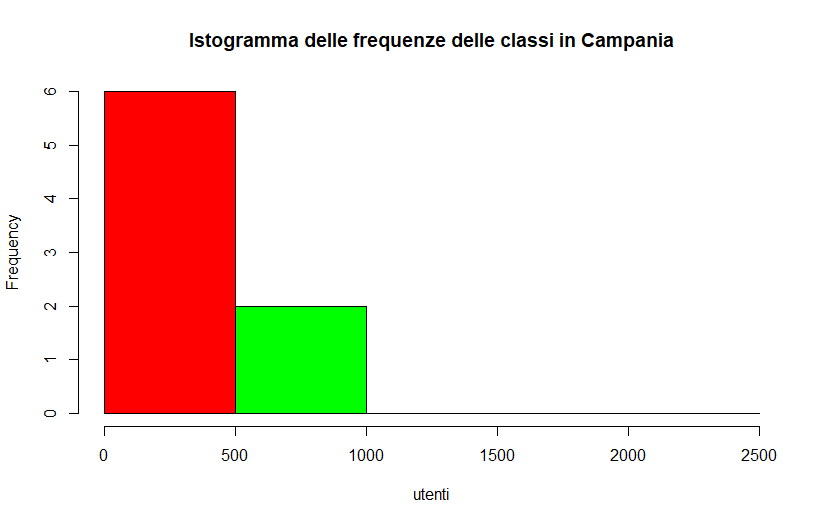
> summary(mediavittimeitalia)

Min. 1st Qu. Median Mean 3rd Qu. Max.

98.77 121.23 133.80 161.27 190.48 262.41

Per individuare la moda bisogna considerare le frequenze dei dati considerando la suddivisione nelle seguenti classi: C1 = [0, 500), C2 = [500, 1000), C3 = [1000, 1500) C4 = [1500, 2000), C5 = [2000, 2500].

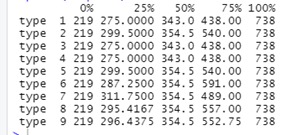
Sia per il primo grafico, che per il secondo, la classe modale è la prima.



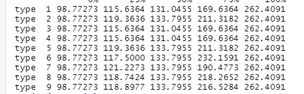
**Quartili con i differenti algoritmi di R**

Poiché ci sono diversi tipi di algoritmi in R per calcolare i quartili, si è deciso di mostrare i risultati anche con diversi algoritmi.

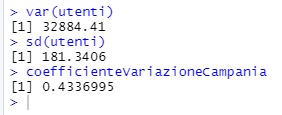
> print(mediavittime)

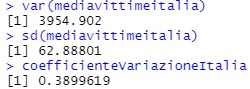


> print(campania)



**Indici di dispersione**





## Forma della distribuzione di frequenze

In questo paragrafo verranno descritti gli indici statistici che permettono di analizzare la forma della distribuzione di frequenze misurando se essa presenta asimmetrie (positive o negative) o se essa è più o meno piccata rispetto ad una distribuzione di frequenze normale standard. Prima di definire tali indici è utile introdurre il concetto di momento campionario e di momento centrato.

Assegnato un insieme di dati numerici x1 , x2 ,…, xn , si definisce **momento campionario** di ordine j la quantità:

Assegnato un insieme di dati numerici x1 , x2 ,…, xn , si definisce **momento campionario centrato** attorno alla media di ordine j la quantità:

La skewness campionaria permette di misurare la simmetria di una distribuzione di frequenze. Assegnato un insieme di dati numerici x1 , x2 ,…, xn , si definisce **skewness campionaria** il valore:

Se la distribuzione è simmetrica il valore γ1 è nullo, γ1 > 0 se la distribuzione ha un’asimmetria positiva (ovvero una coda a destra più allungata), γ1 < 0 se la distribuzione ha un’asimmetria negativa (ovvero una coda a sinistra più allungata).

Il codice per calcolare la skewness campionaria in R è:

skw <-function (x){

n<-length (x)

m2 <-(n -1) \*var (x)/n

m3 <- (sum ( (x- mean(x))^3) )/n

m3/(m2 ^1.5)

}

La **curtosi campionaria** è un indice che permette di misurare la densità dei dati intorno alla media.

Essa si calcola con la seguente equazione:

Dove è l’indice di Pearson e sono rispettivamente il momento centrato campionario di ordine 2 ed ordine 4.

Da notare anche che è indipendente dall’unità di misura dei dati.

Gli indici permettono di confrontare la dei dati con una densità di probabilità normale standard

* Se abbiamo una distribuzione di frequenze platicurtica, è quindi più piatta di una normale
* Se abbiamo una distribuzione di frequenze leptocurtica, è quindi più piccata di una normale
* Se abbiamo una distribuzione di frequenze normocurtica, è quindi piatta come una normale

Il codice per calcolare la curtosi in R è:

curt <-function (x){

n <-length (x)

m2 <-(n -1) \*var (x)/n

m4 <- (sum ((x-mean(x))^4) )/n

m4/(m2 ^2) -3

}

### Forma della distribuzione di frequenze Utenti

**Skewness campionaria**. Entrambe le distribuzioni di frequenze hanno un’asimmetria positiva, la distribuzione di frequenza ha quindi una coda più allungata a destra.

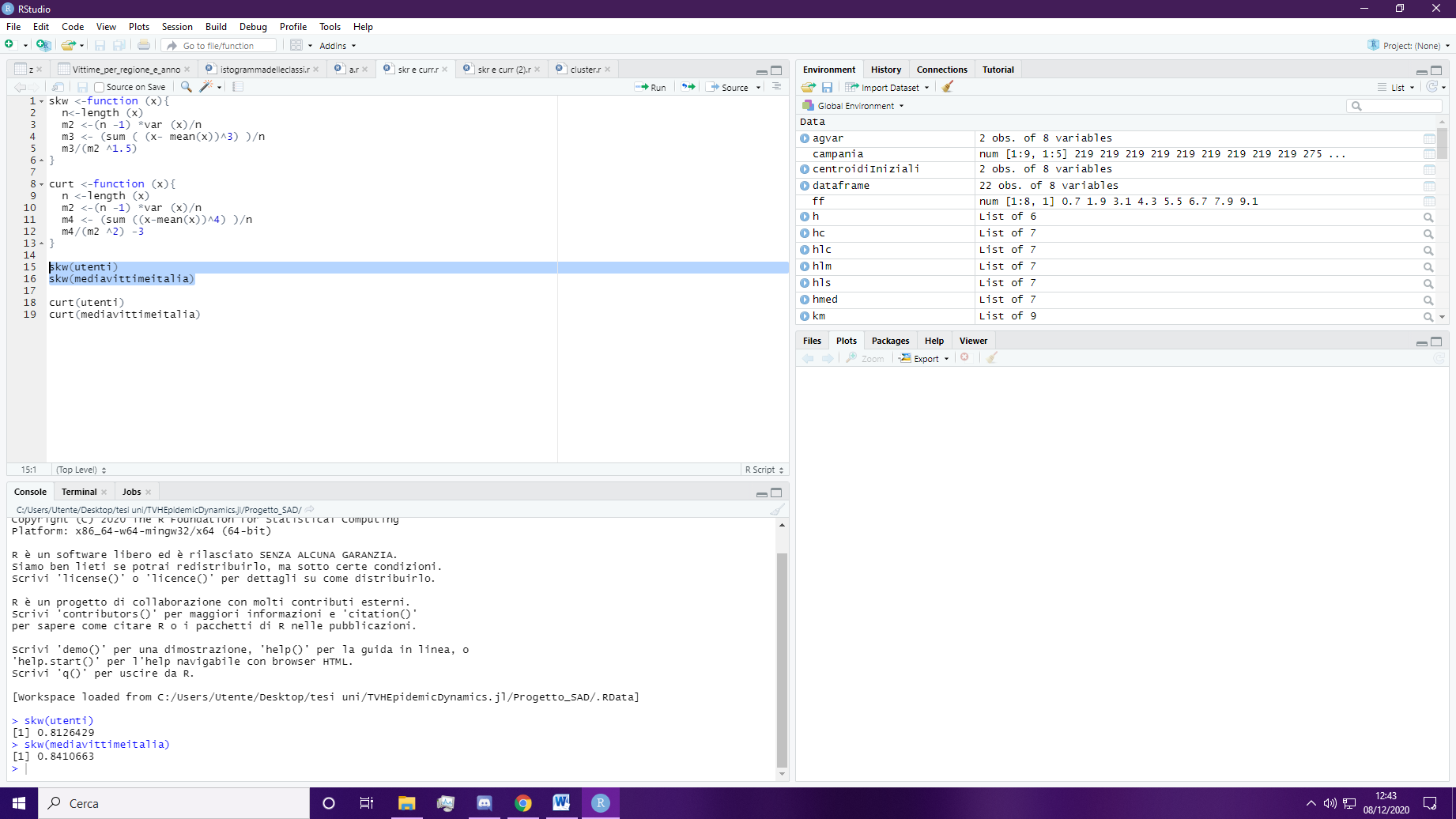


**Curtosi campionaria.** Il valore di entrambe le curtosi campionarie è negativo quindi entrambe le distribuzioni di frequenza sono meno piccate di una distribuzione di frequenze normale standard.

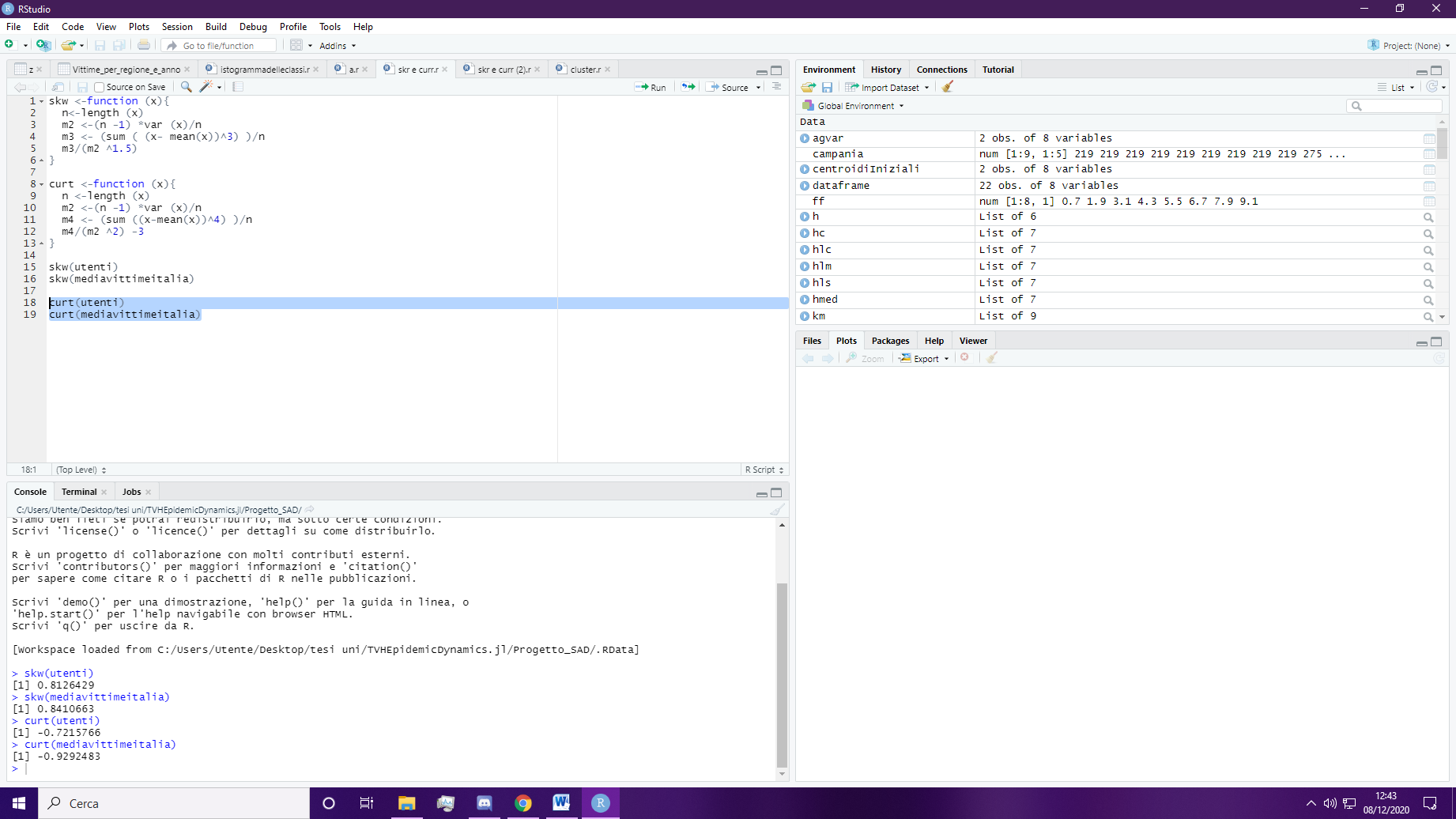


### Forma della distribuzione di frequenze Vittime

**Skewness campionaria**. Per quanto riguarda la skewness campionaria, l’indice è positivo per entrambi i dati, ciò quindi significa che nella distribuzione di frequenze, la coda di destra è più allungata.



**Curtosi campionaria.** Il valore di entrambe le curtosi campionarie è negativo e quindi la distribuzione di frequenze è più piatta di una normale.



# Statistica descrittiva bivariata

In questo capitolo verranno mostrate le analisi di regressione lineare semplice e di regressione lineare multipla calcolando il modello lineare, i residui e il coefficiente di determinazione.

La statistica descrittiva bivariata si occupa dei metodi grafici e statistici atti a descrivere le relazioni che intercorrono tra due variabili X e Y. Un primo passo per indagare l’eventuale dipendenza tra due variabili X e Y consiste nel disegnare il diagramma di dispersione o scatterplot. Per ottenere una misura quantitativa della correlazione tra le variabili si considera la covarianza campionaria.

*Assegnato un campione bivariato (x1,y1),(x2,y2), ...,(xn,yn) di una variabile quantitativa bi-dimensionale (X,Y), siano 𝑥̅ e 𝑦̅ rispettivamente le medie campionarie di x1,x2, ...,xn e di y1,y2, ...,yn. La covarianza campionaria tra le due variabili X e Y è così definita:*

Se le variabili sono correlate positivamente, se le variabili sono correlate negativamente, se le variabili non sono correlate.

Per ottenere una misura quantitativa della correlazione tra le variabili si può anche considerare il coefficiente di correlazione campionario.

*Assegnato un campione bivariato (x1,y1),(x2,y2), ...,(xn,yn) di una variabile quantitativa bidimensionale (X,Y), siano 𝑥̅ e sx la media campionaria e la deviazione standard di x1,x2, ...,xn ed inoltre siano 𝑦̅ e sy la media campionaria e la deviazione standard di y1,y2, ...,yn . Il* ***coefficiente di correlazione campionario*** *tra le due variabili X e Y è cosi definito:*

Il coefficiente di correlazione campionario misura la forza del legame di natura lineare esistente tra due variabili quantitative. In particolare, e il suo valore indica la direzione della retta interpolante.

* : (correlazione perfetta negativa), tutti i punti sono allineati lungo una retta discendente;
* (correlazione negativa), i punti sono posizionati in una nuvola attorno ad una retta interpolante discendente;
* : (nessuna correlazione), i punti sono completamente dispersi in una nuvola che non presenta alcuna evidente direzione di natura lineare;
* : (correlazione positiva), i punti sono posizionati in una nuvola attorno ad una retta interpolante ascendente;
* : (correlazione perfetta positiva), tutti i punti sono allineati lungo una retta ascendente.

## Regressione lineare semplice

Il modello di regressione lineare semplice viene utilizzato per spigare la relazione che esiste tra una variabile dipendente Y e una variabile indipendente X. Il modello di regressione lineare semplice è esprimibile tramite l’equazione di una retta che riesce ad interpolare la nuvola di punti dello scatterplot meglio di tutte le altre possibili rette, dove α è l’intercetta e β è il coefficiente angolare. Se β>0 la retta di regressione è crescente, se β<0 la retta di regressione è discendente, se β=0 la retta è orizzontale. L’intercetta α corrisponde invece al punto di intersezione della retta interpolante con l’asse delle ordinate.

La funzione lm(y˜x) permette di eseguire le analisi di regressione lineare della variabile dipendente y in funzione della variabile indipendente x.

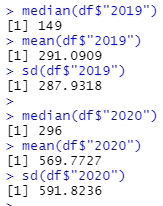
Dopo aver calcolato la retta interpolante, è possibile notare che esistono degli scostamenti tra i valori osservati del campione e i valori stimati attraverso la retta di regressione. Le differenze tra le ordinate dei punti dei valori osservati e le ordinate dei punti dei valori stimati prendono il nome di residui. Se si indica con il valore osservato e con il valore stimato, i **residui** sono così definiti:

Per conoscere quanto la retta di regressione si adatta ai dati considerati si calcola il **coefficiente di determinazione** che viene definito come il rapporto tra la varianza dei valori osservati e la varianza dei valori stimati. Un coefficiente di determinazione prossimo ad 1 indica che tutti i punti tendono ad allinearsi lungo la retta di regressione, mentre un coefficiente di determinazione prossimo a 0 indica una completa incapacità della retta di rappresentare la distribuzione dei dati considerati. Per il modello di regressione lineare semplice il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione.

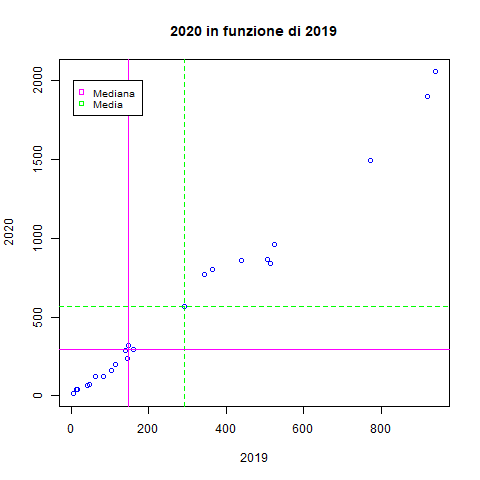
### Regressione lineare semplice Utenti

Le variabili che vengono considerate in quest’analisi sono le colonne della tabella relative ai dati del 2019 e del 2020. In particolare, la variabile indipendente è 2019, quella dipendente è 2020. Si calcolano i valori degli indici statistici mediana, media e deviazione standard dei dati relativi alle variabili considerate.

Si nota che sia mediana, sia media che deviazione standard sono maggiori per la seconda variabile.



Successivamente, si realizza lo scatterplot ponendo sulle ascisse la variabile indipendente 2019 e sulle ordinate la variabile dipendente 2020. Vengono poi tracciate delle linee orizzontali e verticali in corrispondenza delle mediane e delle medie delle due variabili.



Dallo scatterplot si nota che i dati sono posizionati lungo una retta ascendente quindi si può dedurre che esiste una correlazione positiva tra le variabili considerate.

Per vedere se esiste tale correlazione si calcolano la covarianza e la correlazione campionaria. Da questo calcolo si evince che i dati dei due vettori 2019 e 2020 sono positivamente correlati essendo la covarianza positiva. Inoltre, il coefficiente di correlazione è uguale a 0.9923597 che è prossimo ad 1, quindi come indicato dallo scatterplot esiste una forte correlazione lineare tra i dati del 2019 e i dati del 2020.



Il seguente codice permette di realizzare lo scatterplot relativo ai dati del 2019 e del 2020 con la retta interpolante stimata.

png("grafici/bivariata/scatterPlotUtenti2020\_2019RettaRegressione.png")

plot(df$"2019", df$"2020", main="Retta di regressione 2020 in funzione di 2019", col="blue",

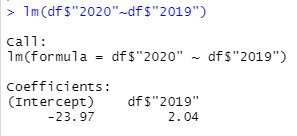
xlab="2019", ylab="2020")

abline(lm(df$"2020"~df$"2019"), col="magenta")

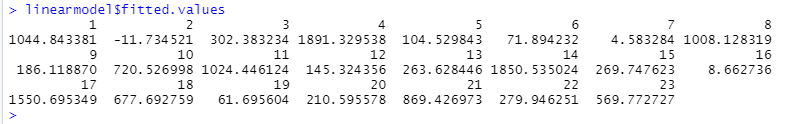
dev.off()



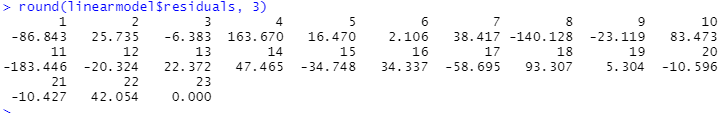
Il seguente codice permette di ottenere il modello di regressione lineare per le due variabili. In particolare, l’intercetta α vale -23.97, mentre il coefficiente angolare β vale 2.04. Siccome il coefficiente angolare ha segno positivo, la retta è ascendente. L’equazione della retta interpolante risulta:



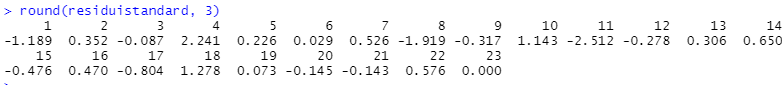
Il codice seguente permette di visualizzare i valori stimati.



Il seguente codice permette di visualizzare i residui, ossia di quanto i valori osservati si discostano dai valori stimati.



Valore dei residui standardizzati rispetto alla deviazione standard. Si può osservare che i valori sono molto piccoli.



Le seguenti linee di codice mostrano i valori della mediana, della varianza e della deviazione standard dei residui.





Il seguente codice crea il grafico mostrato in basso che rappresenta la retta di regressione del 2020 in funzione del 2019. Inoltre, vengono rappresentati dei segmenti di colore verde che congiungono le ordinate dei valori osservati con le ordinate dei valori stimati.

png("grafici/bivariata/scatterPlotUtenti2020\_2019RettaRegressioneResidui.png")

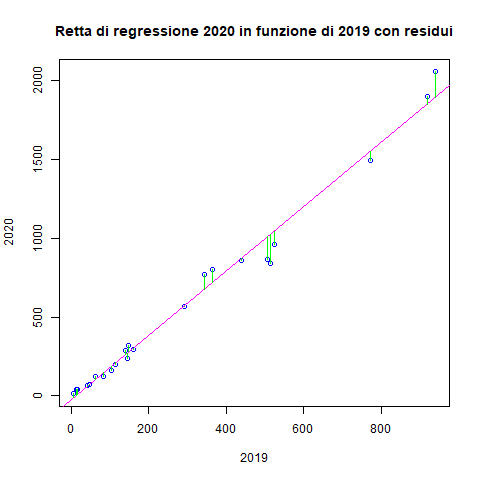
plot(df$"2019", df$"2020", main="Retta di regressione 2020 in funzione di 2019 con residui", col="blue",

xlab="2019", ylab="2020")

abline(linearmodel, col="magenta")

segments (df$"2019", linearmodel$fitted.values, df$"2019", df$"2020" ,col="green")

dev.off()



Per valutare quanto la retta di regressione si adatta ai dati si calcola il coefficiente di determinazione che si calcola effettuando il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati. In questo caso il coefficiente di correlazione vale 0.9848. Siccome è prossimo ad 1, significa che la retta descrive bene i dati considerati, infatti anche dai grafici visti precedentemente si nota che gli scostamenti dalla retta sono molto piccoli.



### Regressione lineare semplice Vittime

In questa sezione verranno utilizzate le colonne del dataset delle vittime del 2019 e del 2020, con la colonna del 2020 come variabile dipendente e quella del 2019 come variabile indipendente.

Si calcolano i valori degli indici statistici mediana, media e deviazione standard per le due colonne.

Sia mediana, che media e deviazione standard sono maggiori per la colonna relativa al 2020.

> median(dataframe$"2019")

[1] 61.5

> mean(dataframe$"2019")

[1] 131.0455

> sd(dataframe$"2019")

[1] 134.5527

> print("cambio anno")

[1] "cambio anno"

> median(dataframe$"2020")

[1] 130.5

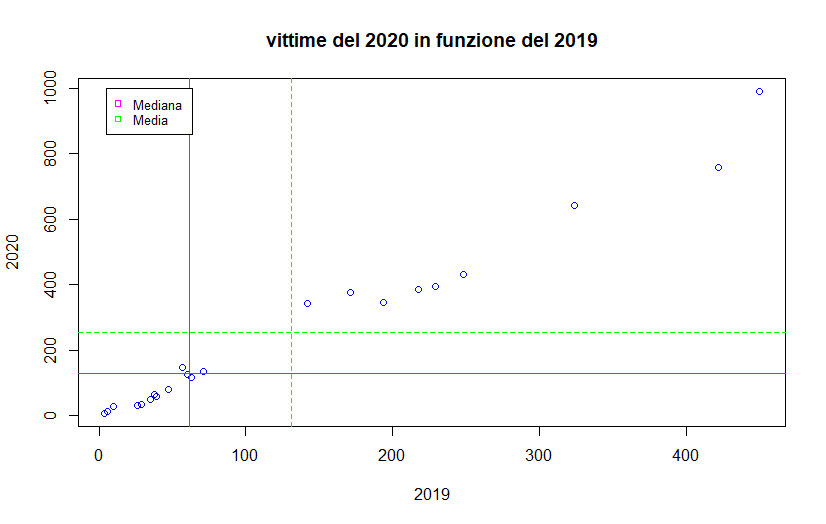
> mean(dataframe$"2020")

[1] 253

> sd(dataframe$"2020")

[1] 269.0815

Successivamente, si realizza lo scatterplot ponendo sulle ascisse la variabile indipendente 2019 e sulle ordinate la variabile dipendente 2020. Vengono poi tracciate delle linee orizzontali e verticali in corrispondenza delle mediane e delle medie delle due variabili.



Dallo scatterplot si può notare come tutti i dati siano posizionati lungo una retta ascendente quindi si può dedurre che esiste una correlazione positiva tra le variabili considerate.

Per vedere se ciò è matematicamente provato, bisogna controllare la covarianza e la correlazione campionaria: la covarianza deve essere positiva e la correlazione deve essere prossima ad 1.

> cov(dataframe$"2019", dataframe$"2020")

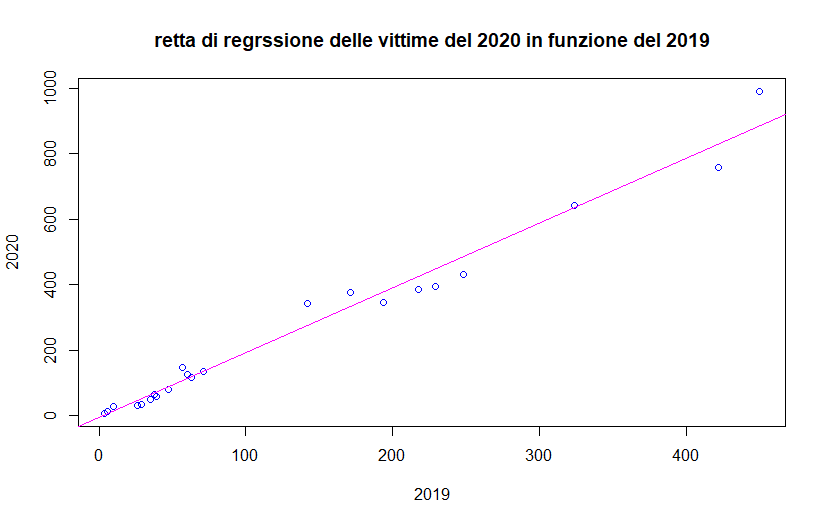
[1] 35798.62

> cor(dataframe$"2019",dataframe$"2020")

[1] 0.9887581

Entrambe le condizioni sono soddisfatte. Le due variabili sono linearmente correlate poiché la correlazione è prossima ad 1, ed è positiva poiché la covarianza è positiva.

Il seguente grafico mostra lo scatterplot relativo ai dati del 2019 e del 2020 con la retta interpolante stimata.



Il seguente codice permette di ottenere il modello di regressione lineare per le due variabili. In particolare, l’intercetta vale -6.122, mentre il coefficiente angolare vale 1.977. Siccome il coefficiente angolare è positivo, la retta è ascendente. L’equazione della retta risulta quindi:

Call:

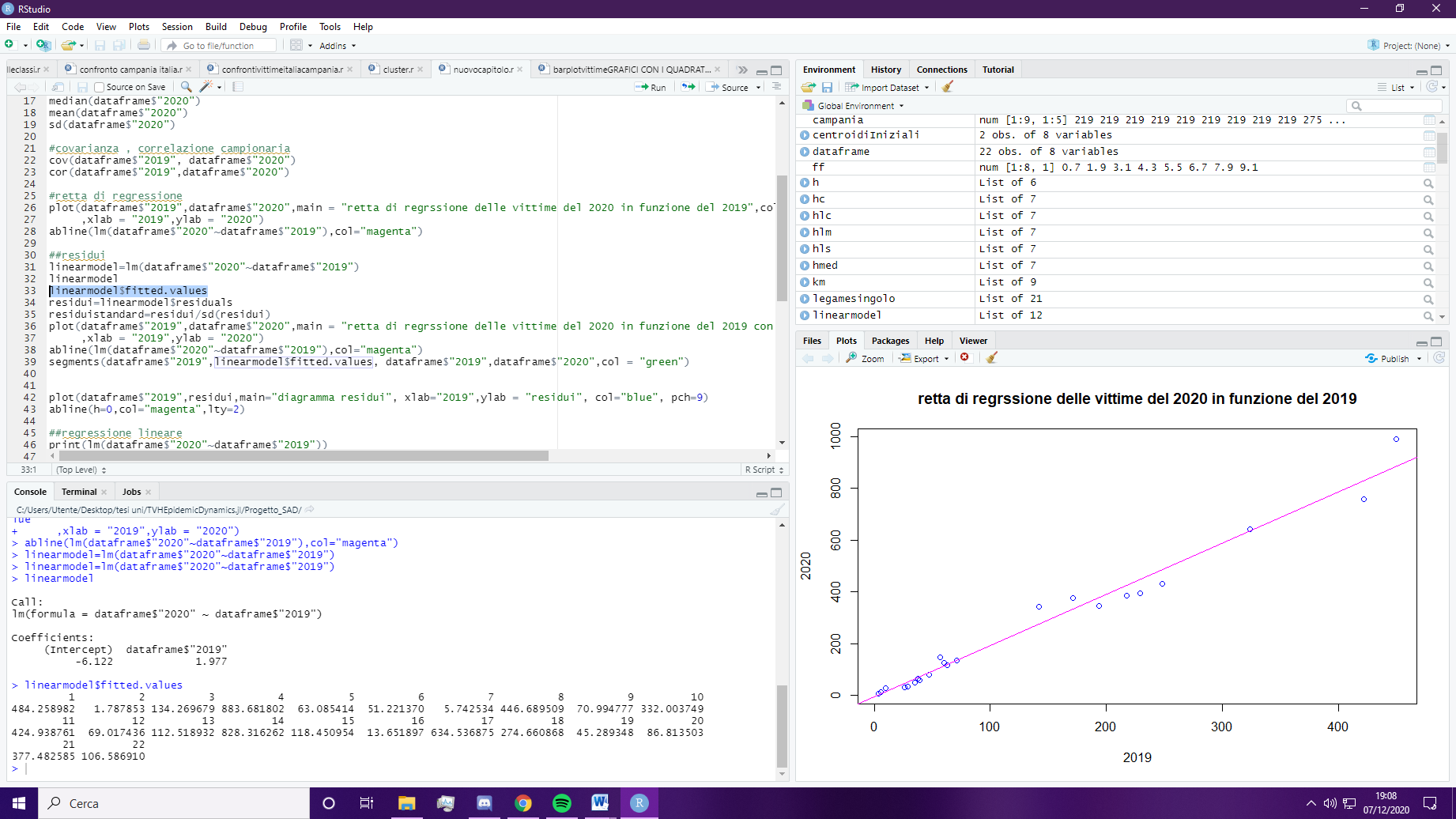
lm(formula = dataframe$"2020" ~ dataframe$"2019")

Coefficients:

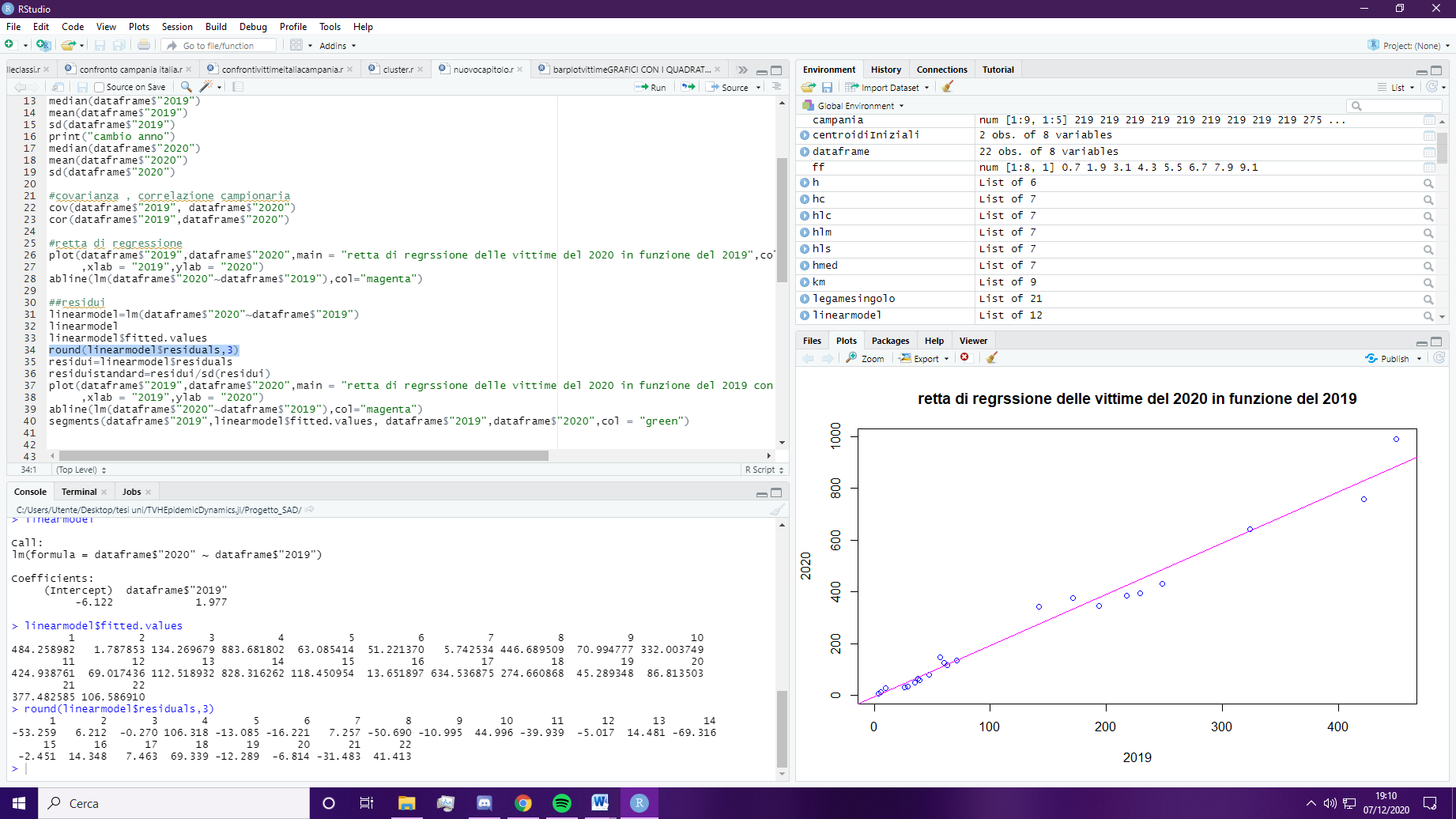
(Intercept) dataframe$"2019"

-6.122 1.977

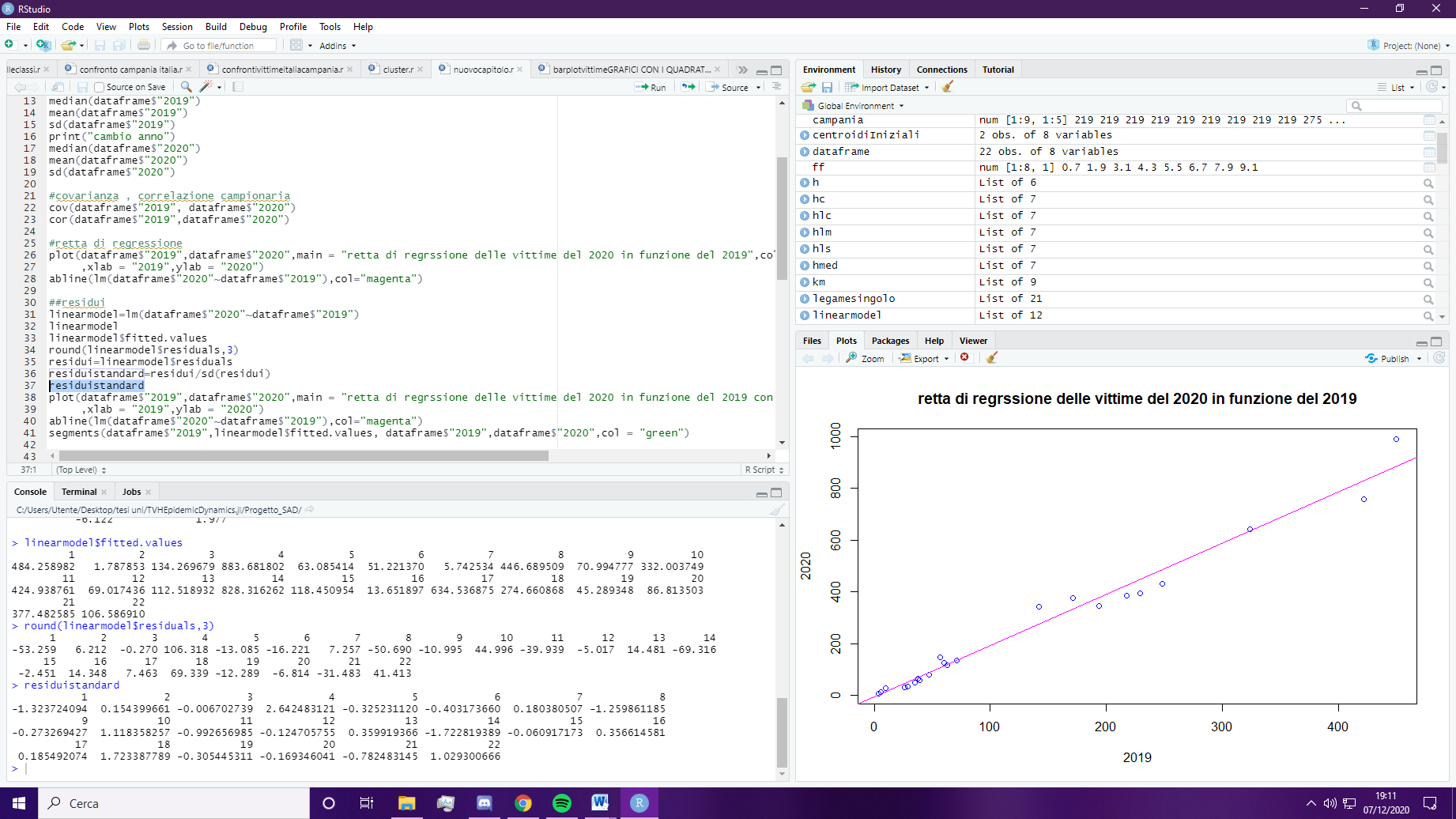
Il codice seguente permette di visualizzare i valori stimati.



Il seguente codice permette di visualizzare i residui, ossia di quanto i valori osservati si discostano dai valori stimati.

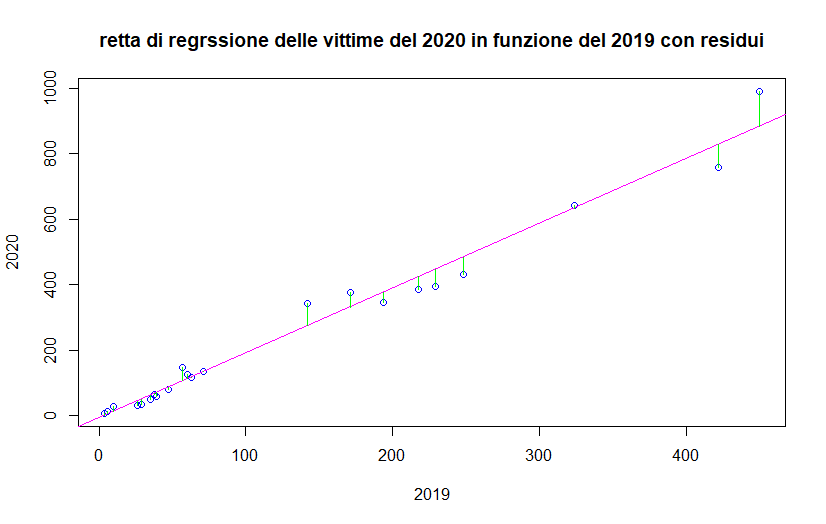


Valore dei residui standardizzati rispetto alla deviazione standard. Si può osservare che i valori sono molto piccoli.



Le seguenti linee di codice mostrano i valori della mediana, della varianza e della deviazione standard dei residui.





Per valutare quanto la retta di regressione si adatta ai dati si calcola il coefficiente di determinazione che si calcola effettuando il rapporto tra la varianza dei valori stimati tramite la retta di regressione e la varianza dei valori osservati. In questo caso il coefficiente di correlazione vale 0.9776425. Siccome è prossimo ad 1, significa che la retta descrive bene i dati considerati, infatti anche dai grafici visti precedentemente si nota che gli scostamenti dalla retta sono molto piccoli.



## Regressione lineare multipla

Il modello di regressione lineare multipla viene utilizzato per spiegare la relazione tra una variabile quantitativa Y detta variabile dipendente e le variabili quantitative indipendenti X1, X2, …, Xp.

Il modello di regressione lineare multipla con p variabili indipendenti è esprimibile attraverso l’equazione:

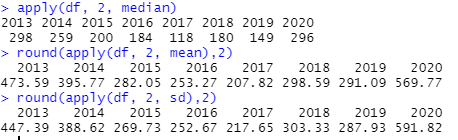
Dove:

* è l’intercetta, ossia il valore di Y quando X1=X2= …= Xp=0;
* sono i regressori. In particolare, rappresenta l’inclinazione di Y rispetto alla variabile X1 tenendo costanti le variabili X2, X3, …,Xp, …, rappresenta l’inclinazione di Y rispetto alla variabile Xp tenendo costanti le variabili X1, X2, …,Xp-1.

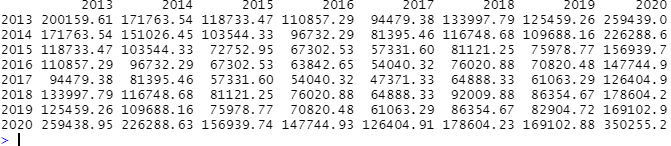
### Regressione lineare multipla Utenti

Si utilizza il modello di regressione lineare multipla per spiegare la relazione le variabili indipendenti: 2013, 2014, 2015, 2016, 2017, 2018, 2019 e la variabile dipendente: 2020

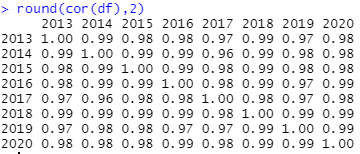
Valore degli indici di posizione e di dispersione (mediana, media e deviazione standard) relativi alle variabili:



Di seguito viene mostrata la matrice delle covarianze che contiene sulla diagonale principale la varianza delle singole colonne del dataframe, mentre gli altri elementi rappresentano le covarianze tra le coppie di variabili.



Di seguito viene mostrata la matrice delle correlazioni che contiene tutte le correlazioni lineari tra le coppie di variabili, ossia misura la forza del legame di natura lineare esistente tra tutte le coppie di variabili quantitative. La matrice delle correlazioni contiene 1 sulla diagonale principale.



Si nota che esiste una forte correlazione lineare tra tutte le variabili considerate.

Il seguente grafico visualizza in un’unica finestra tutti gli scatterplot ottenuti mettendo in relazione le varie coppie di variabili. Da tale grafico si può dedurre che le variabili sono altamente correlate e si intuisce che avranno un coefficiente di correlazione quasi pari ad 1.



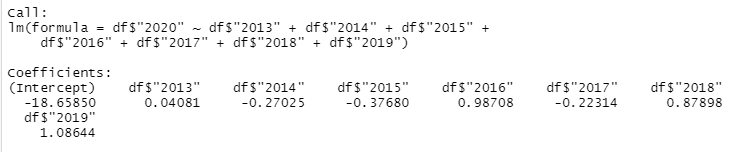
Il grafico precedente è ottenuto mediante il codice:

png("grafici/bivariata/scatterPlotUtentiCoppieVariabili.png")

pairs(df, main="Scatterplot per le coppie di variabili", col="blue")

dev.off()

Utilizzando il modello di regressione lineare multipla si ottiene:

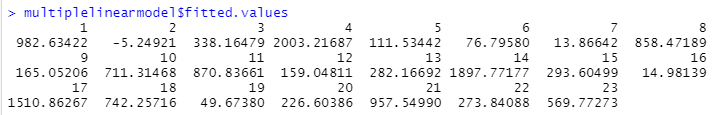


Da cui si ricava che l’intercetta è -18.65850 e i regressori sono: 0.04081, -0.27025, -0.37680, 0.98708, -0.22314, 0.87898, 1.08644.

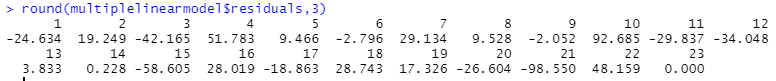
I segni dei regressori β1, β4, β6, β7 sono positivi: questo indica che all’aumentare del numero di utenti nel 2013, 2016, 2018 e 2019 aumenta il numero di utenti nel 2020. Mentre i regressori β2, β3 , β5 sono negativi quindi all’aumentare del numero di utenti nel 2014, 2015, 2017 diminuisce il numero di utenti nel 2020.

Il regressore di 2013 è prossimo allo zero, questo indica che il numero di utenti nel 2013 non incide in maniera significativa il numero di utenti nel 2020.

Valori stimati rispetto al modello di regressione multipla.



Residui dei valori osservati rispetto ai valori stimati.



Valori dei residui standardizzati rispetto alla deviazione standard.

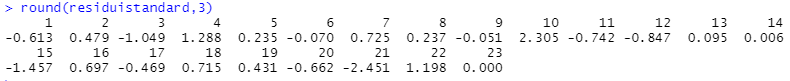
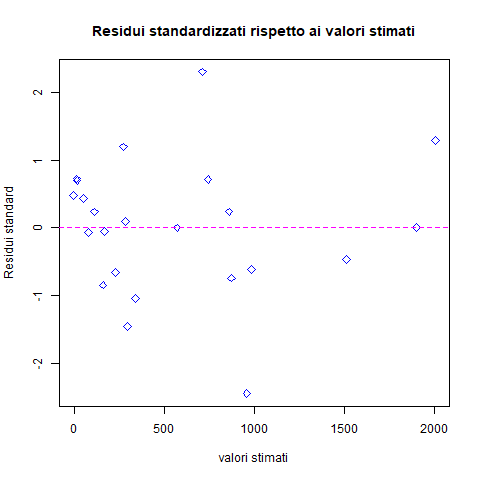


Grafico che mostra i residui standardizzati in funzione dei valori stimati.



I punti indicano dove si collocano i residui rispetto ai valori stimati. Non si evidenzia nessuna tendenza particolare rispetto alla retta orizzontale che rappresenta la media dei residui (0).

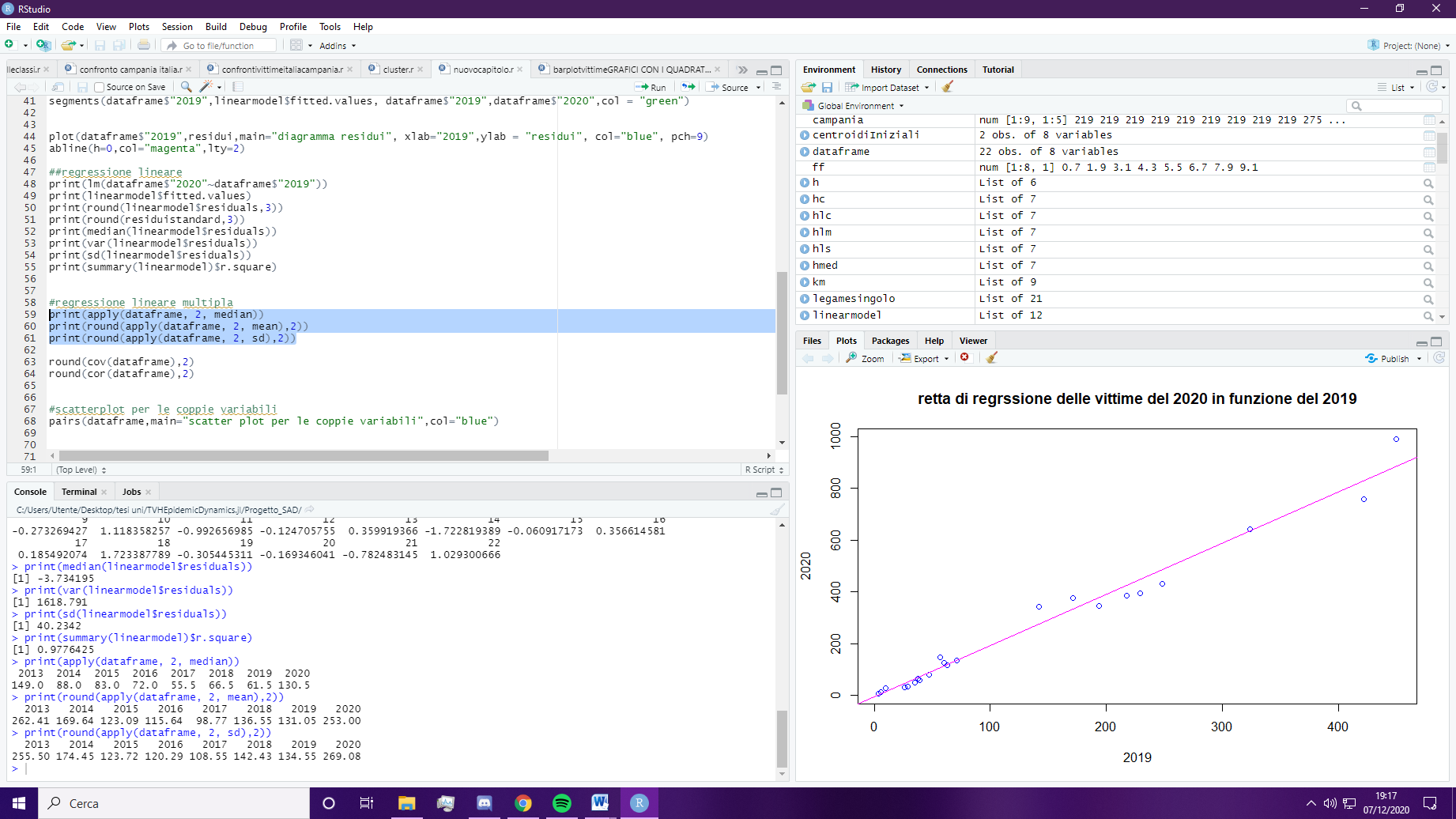
Anche in questo caso il coefficiente di determinazione è prossimo ad 1, infatti vale 0.9954. Il modello di regressione lineare multipla descrive bene i dati considerati.



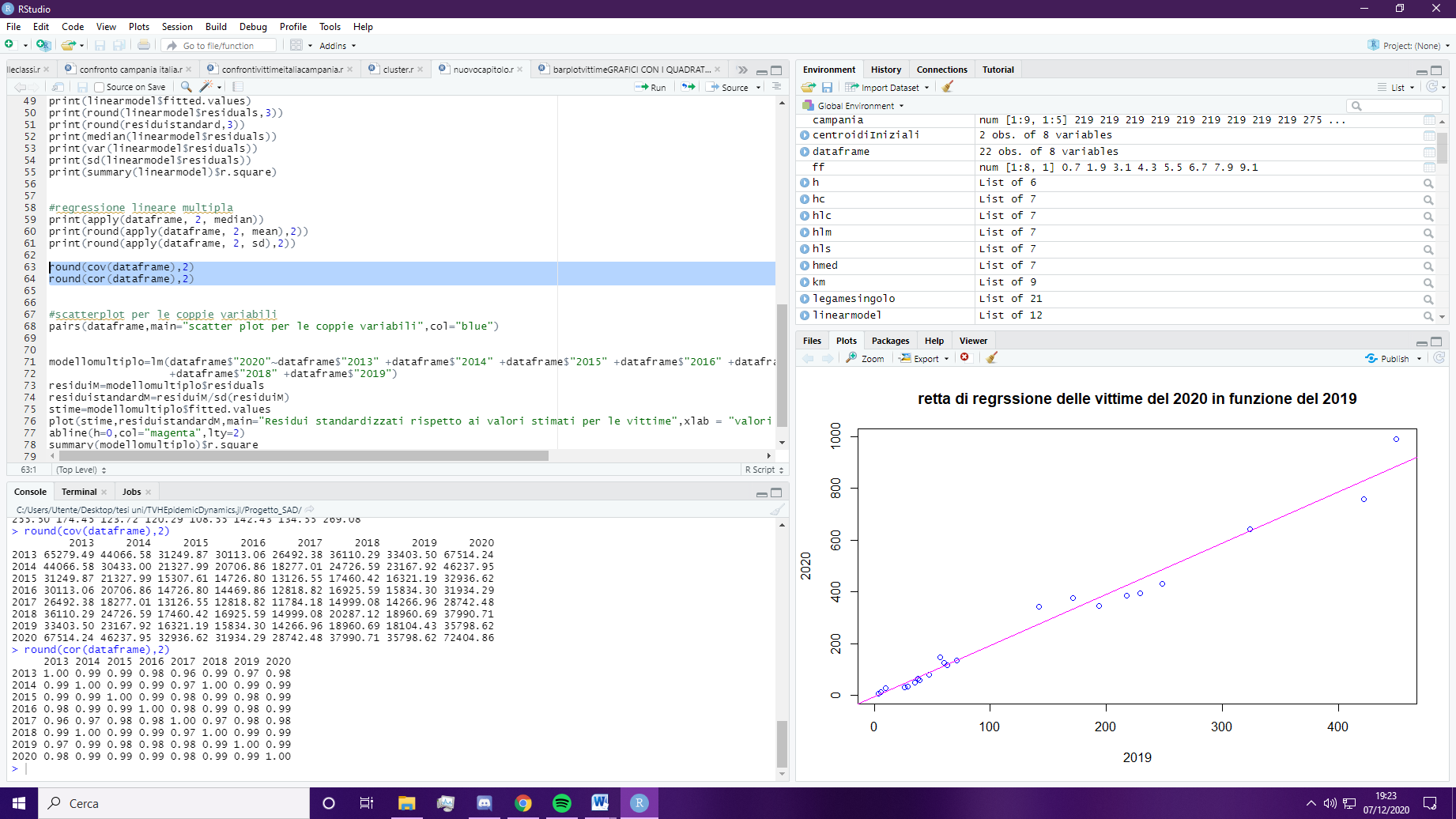
### Regressione lineare multipla Vittime

Si utilizza il modello di regressione lineare multipla per spiegare la relazione le variabili indipendenti: 2013, 2014, 2015, 2016, 2017, 2018, 2019 e la variabile dipendente: 2020

Valore degli indici di posizione e di dispersione (mediana, media e deviazione standard) relativi alle variabili:

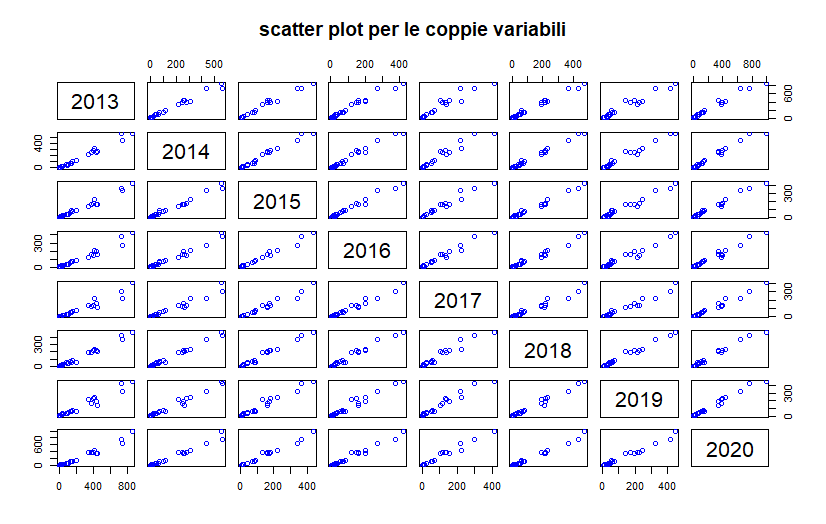


Matrice delle covarianze e sotto la matrice delle correlazioni che contiene tutte le correlazioni lineari tra le coppie di variabili:

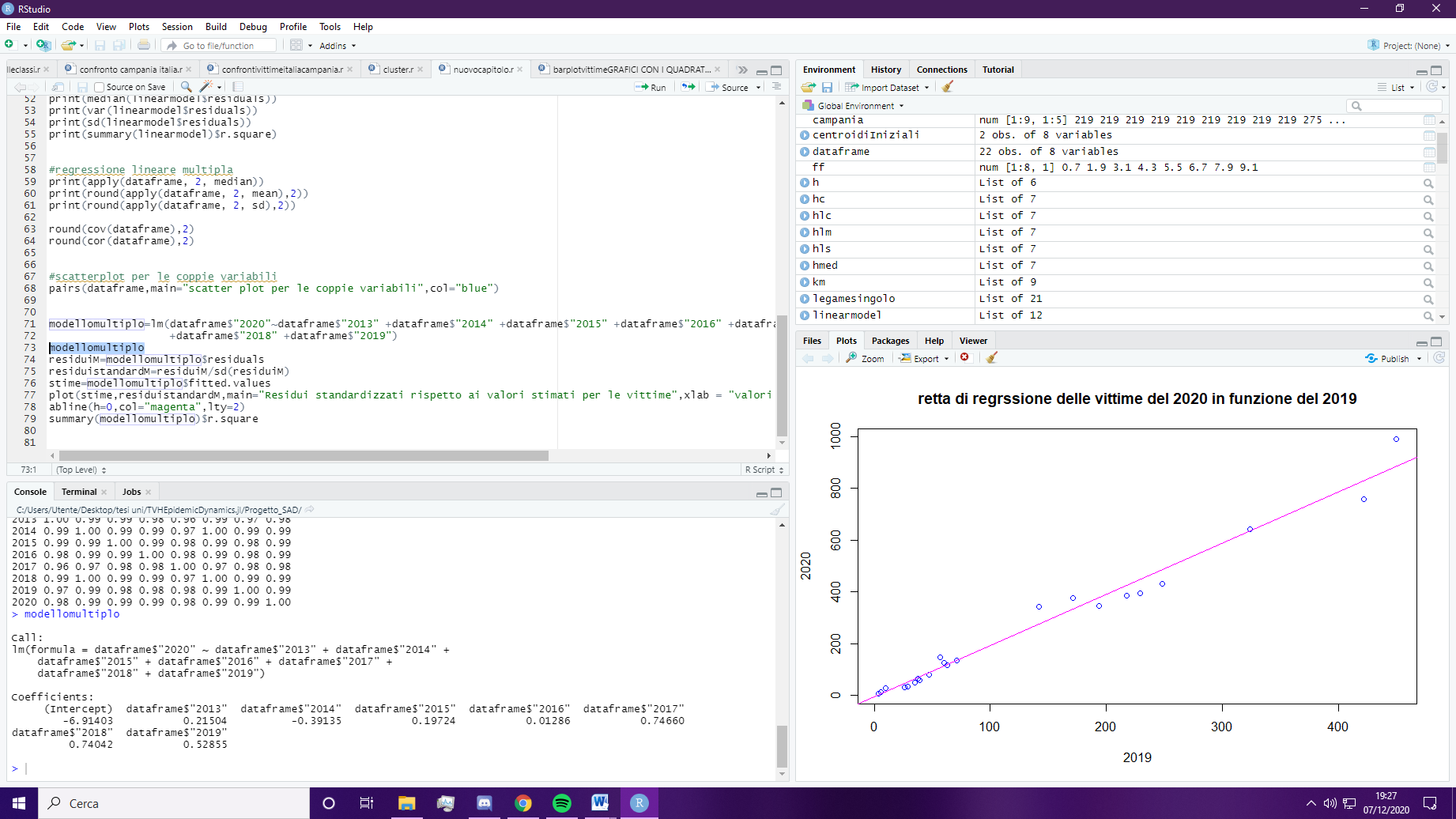


Si nota che esiste una forte correlazione lineare tra tutte le variabili considerate.

Il seguente grafico visualizza in un’unica finestra tutti gli scatterplot ottenuti mettendo in relazione le varie coppie di variabili. Da tale grafico si può dedurre che le variabili sono altamente correlate e si intuisce che avranno un coefficiente di correlazione quasi pari ad 1.



Utilizzando il modello di regressione lineare multipla si ottiene:

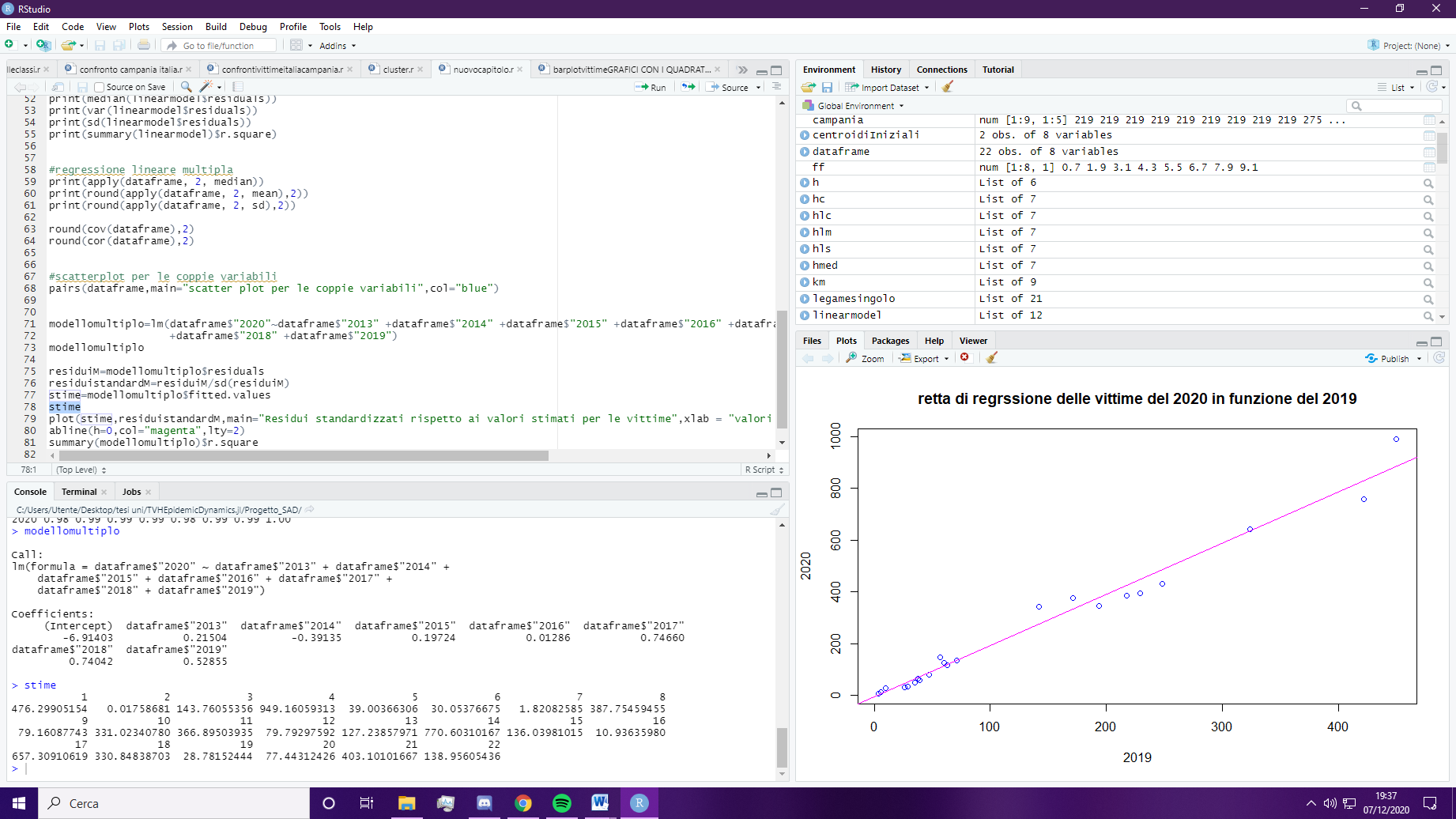


Da cui si ricava che l’intercetta è -6.91403 e i regressori sono: 0.21504, -0.39135, 0.19724, 0.01286, 0.74660, 0.74042, 0.52855.

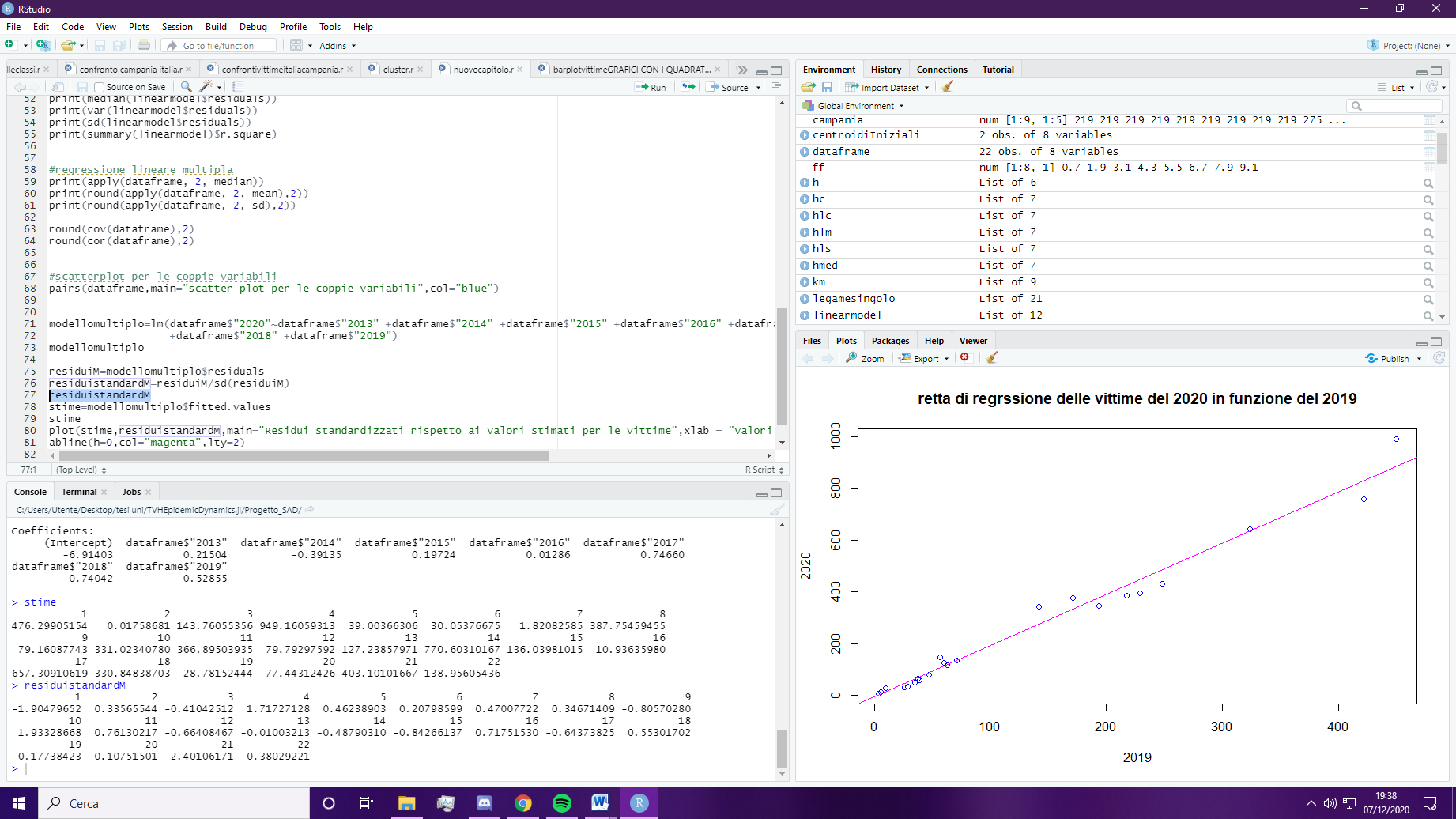
I segni dei regressori β1, β3, β4, β5, β6, β7, β8 sono positivi: questo indica che all’aumentare del numero di utenti nel 2013, 2015, 2016, 2017,2018,2019 aumenta il numero di utenti nel 2020. Mentre il regressore β2 è negativo quindi all’aumentare del numero di utenti nel 2014 diminuisce il numero di utenti nel 2020.

Il regressore del 2016 è prossimo allo zero, ciò vuol dire che l’aumento di vittime nel 2016 non incide molto sulle vittime del 2020.

Valori stimati rispetto al modello di regressione multipla:



Valori dei residui standardizzati rispetto alla deviazione standard.



Residui dei valori osservati rispetto ai valori stimati.

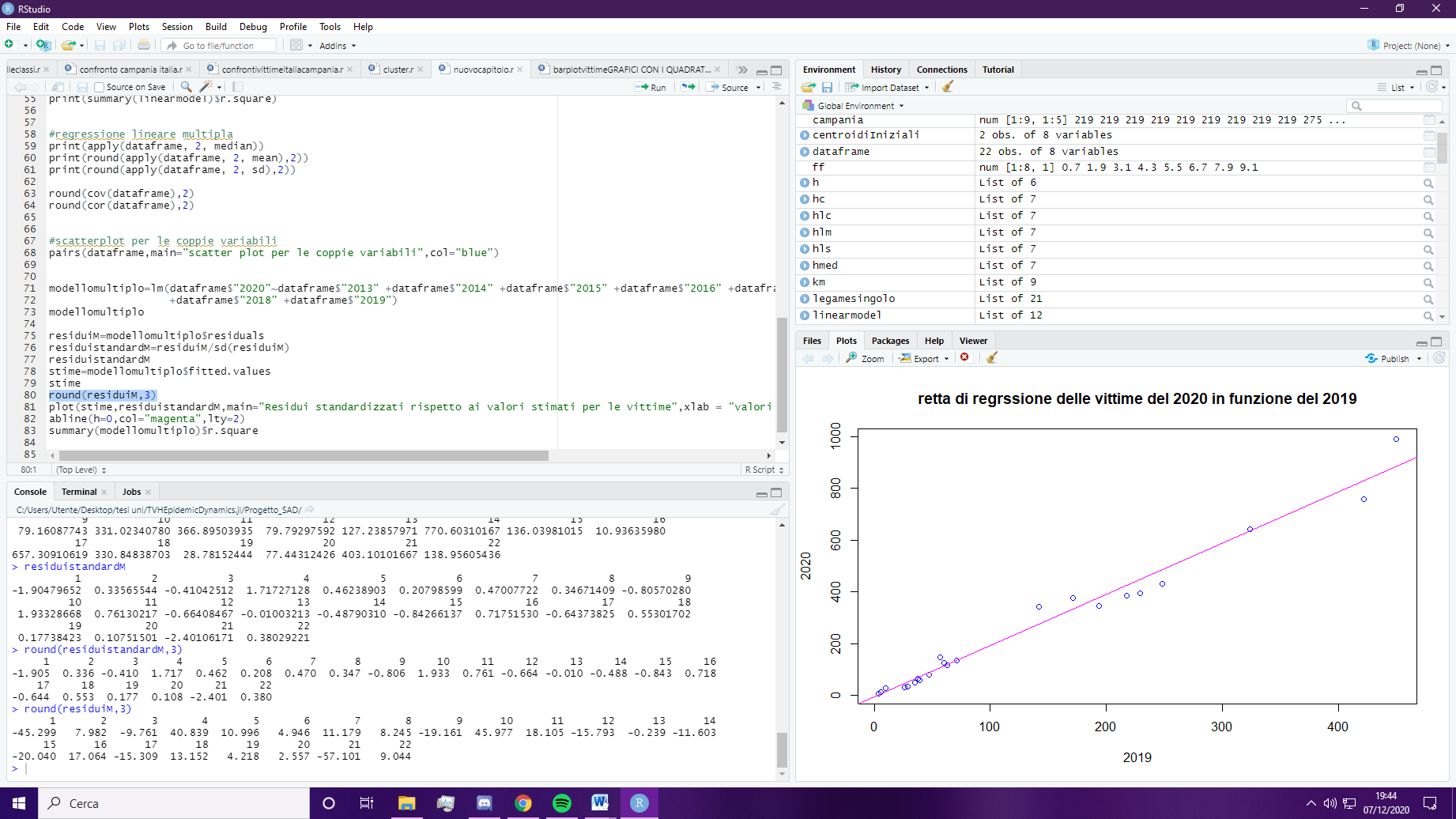
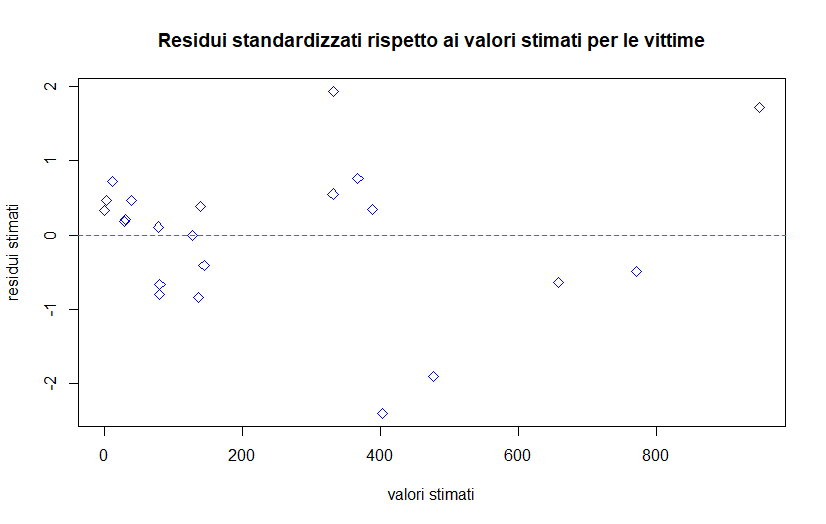
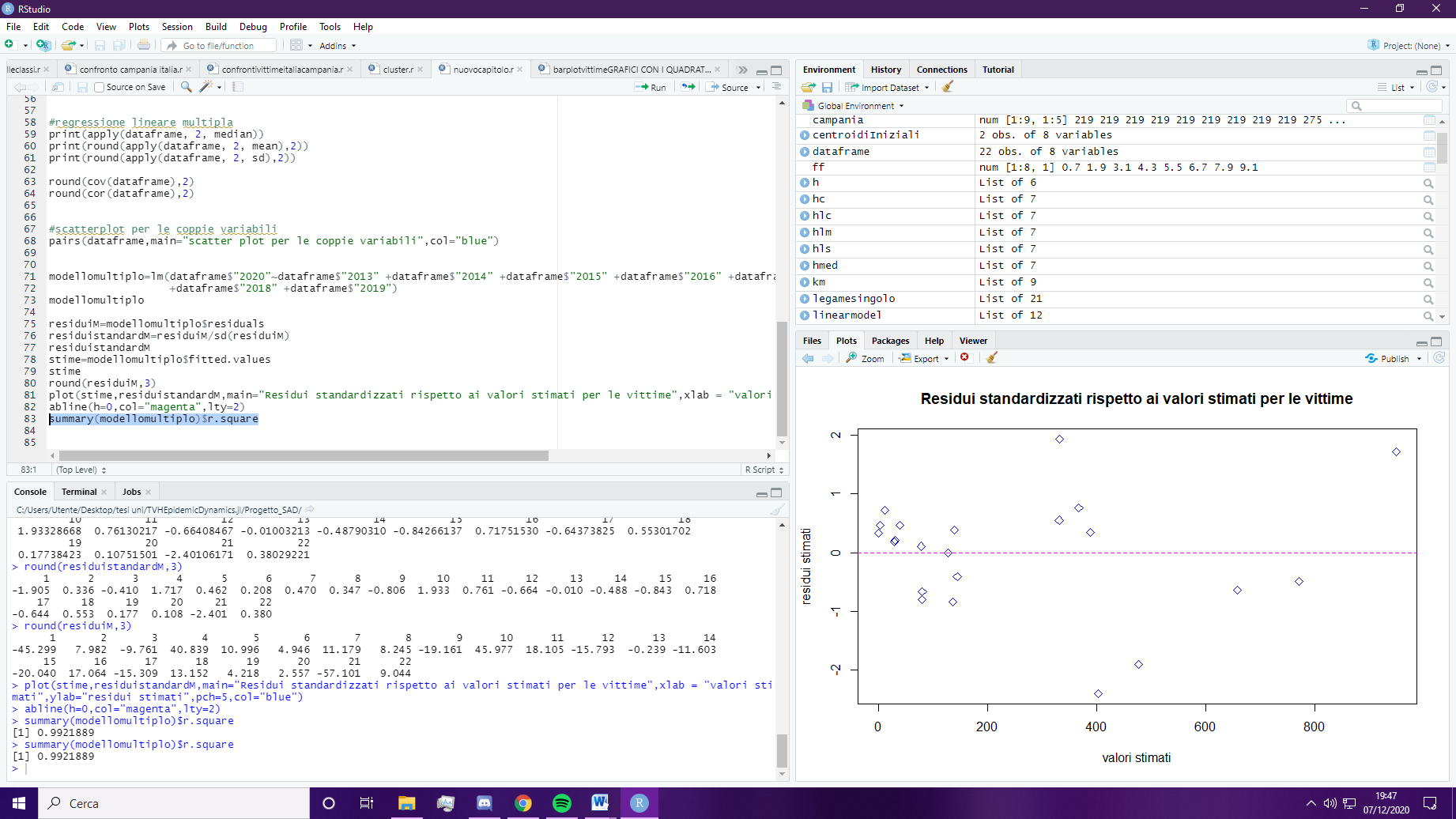


Grafico che mostra i residui standardizzati in funzione dei valori stimati.



I punti indicano dove si collocano i residui rispetto ai valori stimati. Non si evidenzia nessuna tendenza particolare rispetto alla retta orizzontale che rappresenta la media dei residui (0).

Anche in questo caso il coefficiente di determinazione è prossimo ad 1, infatti vale 0.9921889. Il modello di regressione lineare multipla descrive bene i dati considerati



# Analisi dei cluster

L’**analisi dei cluster** è una tecnica matematica usata in informatica e altre discipline, essa si basa sul considerare diversi tipi di dati (numerici, persone, misure) ed unirli in gruppi che contengono tutti elementi che hanno somiglianze tra di loro. La creazione dei cluster può essere effettuata in diversi metodi, ma tutte le tecniche hanno in comune lo scopo di rendere quanto più possibili omogenei gli elementi all’interno di un gruppo e rendere quanto il più eterogenei i gruppi così che il grado di associazione sia alto tra membri dello stesso gruppo e basso tra membri di gruppi diversi.

Le tecniche di raggruppamento tendono ad unire quei dati che sono tra di loro simili e svolgono questo lavoro basandosi sul concetto che ogni elemento di un certo insieme di dati ha delle caratteristiche osservabili, possono essere il colore degli occhi per le persone, o possono essere le denunce al numero verde 1522 fatte di anno in anno per una regione ed è per questo che si usano le funzioni distanza tra i vettori delle caratteristiche che servono a calcolare le misure metriche di somiglianza.

La misura di distanza utilizzata nel progetto è la **metrica Euclidea** così definita:

Dove è il valore della k-esima caratteristica dell’individuo I.

Ottenuta la distanza dei vari elementi è necessario raggrupparli in cluster. Esistono due tipologie per creare i cluster.

* **Metodi gerarchici**: mirano a costruire gerarchie di cluster; si dividono in due tipologie di approcci diversi: L’approccio agglomerativo è un approccio “bottom-up”, si parte dall’inserire ogni elemento in un singolo cluster e si procede ad accorparli a due a due; l’approccio divisivo è un approccio “top-down” che da un singolo cluster che comprende tutti gli elementi viene diviso in tanti sotto cluster. Tutti i metodi gerarchici producono una struttura ad albero chiamata “dendogramma”.
* **Metodi non gerarchici**: permettono di riposizionare elementi di un cluster qualora venga notato che un elemento piazzato in cluster conviene spostarlo in un altro, di questo metodo fa parte l’algoritmo k-means.

Tra i metodi gerarchici ci sono:

* **Metodo del legame singolo**: Presa la matrice delle distanze, si parte dalla distanza 0; si trovano gli elementi che hanno la distanza minore, si rimuovono quegli elementi e si cercano i primi elementi che hanno la distanza minore. Se ci sono più coppie di elementi che hanno la distanza minore nella matrice se ne sceglie uno arbitrariamente.
* **Metodo del legame completo**: La distanza tra due gruppi g1 e g2, con n1 e n2 individui, è definita come la massima tra tutte le distanze di n1 e n2, questo metodo privilegia la differenza tra i gruppi piuttosto che l’omogeneità del gruppo stesso.
* **Metodo del legame medio**: nel metodo del legame medio si considera, come distanza tra due gruppi, la media di tutte le distanze calcolate a due a due tra tutti gli elementi dei due gruppi
* **Metodo del centroide**: la distanza tra i gruppi g1 e g2 è calcolata sulle medie campionarie dei due gruppi. La particolarità di questo metodo è che tende ad avere un effetto gravitazionale: I gruppi più grandi tendono ad assorbire i gruppi più piccoli.
* **Metodo della mediana**: il metodo è simile a quello del centroide, ma non è dipendente dalla numerosità del gruppo. Quando due gruppi si uniscono, il nuovo centroide è calcolato come la semisomma dei due gruppi precedenti.

Tra i metodi non gerarchici, il metodo usato nel progetto è stato “**k-means**”, l’algoritmo funziona in diversi step:

1. Si fissano a priori il numero dei cluster scegliendo però elementi che hanno determinate caratteristiche
2. Si considerano tutti gli elementi e si attribuisce ad ognuno un cluster basandosi sulla distanza minore dal punto di riferimento scelto per ogni cluster
3. Si ricalcolano i centroidi dei k gruppi costituendo il nuovo punto di riferimento per i cluster così ottenuti
4. Si rivalutano le distanze per ogni unità rispetto ai centroidi dei vari cluster. Se un elemento x ha una distanza minore ad un altro centroide rispetto a quello del proprio cluster, si riposiziona l’elemento.
5. Si ricalcolano i centroidi.
6. Si ripete dallo step 4, se si arriva ad un punto in cui non ci sono stati spostamenti tra elementi dei cluster, l’algoritmo si conclude.

## Suddivisione in cluster Utenti

Il seguente codice permette di calcolare la matrice delle distanze euclidee a partire dal data frame Z.

d<-dist(Z, method="euclidean", diag=TRUE, upper=TRUE)

**Metodo del legame singolo**

hls<-hclust(d, method="single")

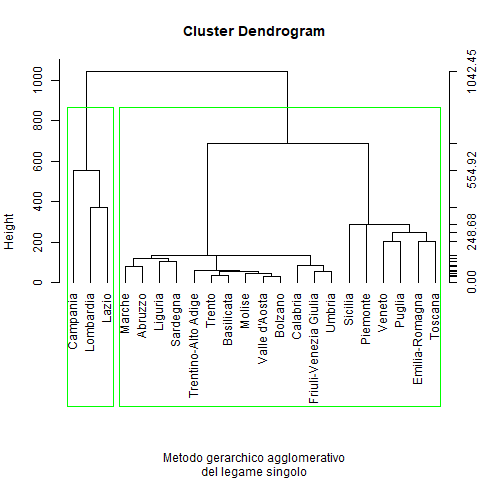
png("grafici/cluster/dendrogrammaUtenti\_LegameSingolo.png")

plot(hls, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame singolo")

rect.hclust(hls, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo del legame medio**

hlm<-hclust(d, method="average")

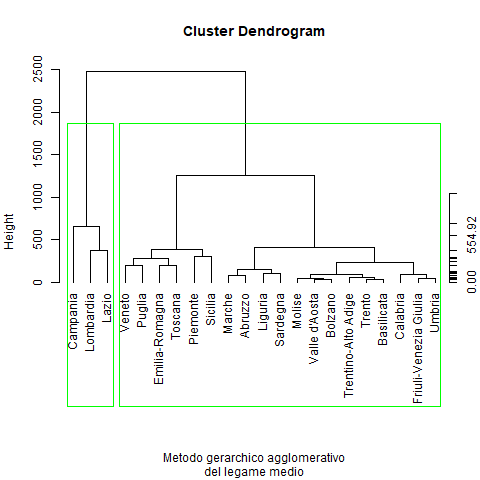
png("grafici/cluster/dendrogrammaUtenti\_LegameMedio.png")

plot(hlm, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame medio")

rect.hclust(hlm, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo del legame completo**

hlc<-hclust(d, method="complete")

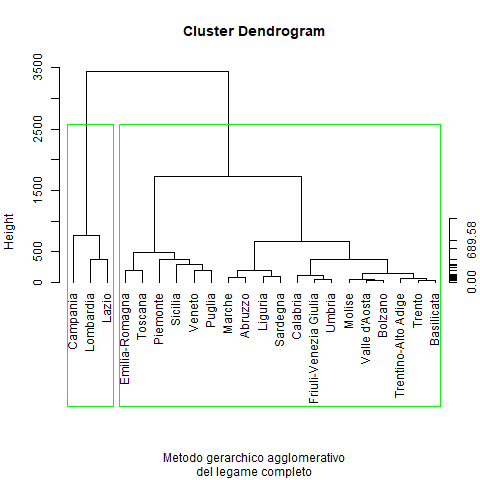
png("grafici/cluster/dendrogrammaUtenti\_LegameCompleto.png")

plot(hlc, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del legame completo")

rect.hclust(hlc, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo del centroide**

d2<-d^2

hc<-hclust(d2, method="centroid")

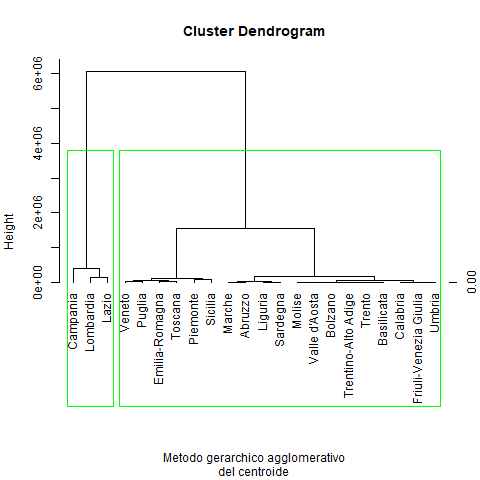
png("grafici/cluster/dendrogrammaUtenti\_MetodoCentroide.png")

plot(hc, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="del centroide")

rect.hclust(hc, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



**Metodo della mediana**

hmed<-hclust(d2, method="median")

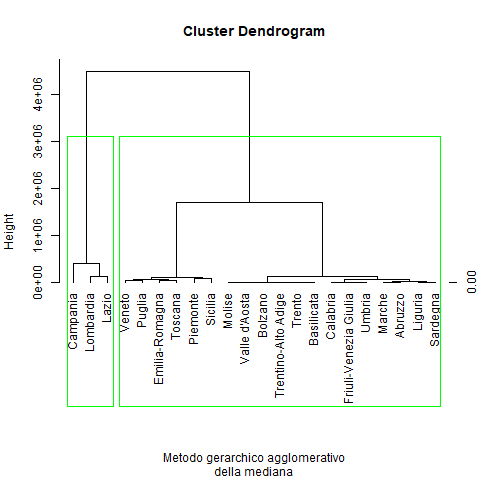
png("grafici/cluster/dendrogrammaUtenti\_MetodoMediana.png")

plot(hmed, hang=-1, xlab="Metodo gerarchico agglomerativo", sub="della mediana")

rect.hclust(hmed, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))

dev.off()



Tutti i metodi gerarchici: legame singolo, legame medio, legame completo, metodo del centroide e metodo della mediana hanno fornito il seguente partizionamento in due cluster.

Primo cluster: 19 individui

Secondo cluster: 3 individui

|  |  |
| --- | --- |
| Cluster 1 | Piemonte, Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Abruzzo, Molise, Puglia, Basilicata, Calabria, Sicilia, Sardegna |
| Cluster 2 | Lombardia, Lazio, Campania |

Si mostra in R il codice per il calcolo delle misure di non omogeneità per i cluster ottenuti con il metodo del legame singolo. Siccome il partizionamento ottenuto è uguale anche per gli altri metodi i risultati saranno uguali.

n<-nrow(Z)

trH<-(n -1)\*sum (apply(Z,2,var))

taglio<-cutree(hls, k=2)

num <-table (taglio)

tagliolist<-list(taglio)

agvar <- aggregate (Z, tagliolist , var)[, -1]

trH1 <-(num [[1]]-1) \* sum (agvar [1, ])

trH2 <-(num [[2]]-1) \* sum (agvar [2, ])

trB<-trH-trH1-trH2

rapportoLegameSingolo<-trB/trH

La misura di non omogeneità totale trH nel data frame considerato risulta essere uguale a 23327101.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 7304986.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 340968.7.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2= 15681146.

Il rapporto =  **0.6722287**.

Il metodo non gerarchico K-means ha fornito il seguente partizionamento in due cluster.

Primo cluster: 9 individui

Secondo cluster: 13 individui

|  |  |
| --- | --- |
| Cluster 1 | Piemonte, Lombardia, Veneto, Emilia-Romagna, Toscana, Lazio, Campania, Puglia, Sicilia |
| Cluster 2 | Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Friuli-Venezia Giulia, Umbria, Marche, Abruzzo, Molise, Basilicata, Calabria, Sardegna |

km <-kmeans (Z, centers=2, iter.max =10, nstart =1)

rapportoKMeans<-km$betweenss/km$totss

La misura di non omogeneità totale trH nel data frame considerato risulta essere uguale a 23327101.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 5812676.7.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 637988.3.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2=16876436.

Il rapporto = **0.7234691**.

La suddivisione in cluster ottenuta con il metodo non gerarchico K-means risulta essere migliore.

## Suddivisione in cluster Vittime

Il seguente codice permette di calcolare la matrice delle distanze euclidee a partire dal dataframe Z.

d=dist(z,method = "euclidean",diag=TRUE,upper=TRUE)

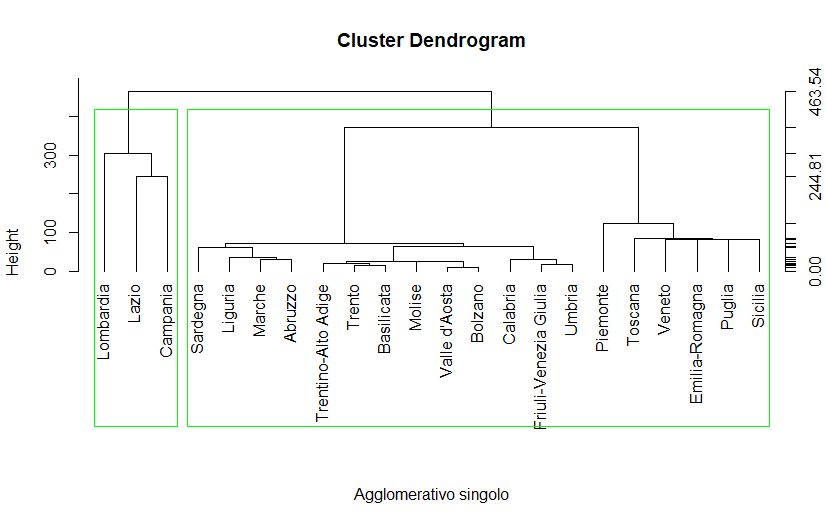
**Metodo del legame singolo**

hls=hclust(d,method = "single")

plot(hls,hang=-1,xlab = "Agglomerativo singolo",sub=" ")

rect.hclust(hls,k=2,border = "green")

axis(side=4,at=round(c(0,hls$height),2))



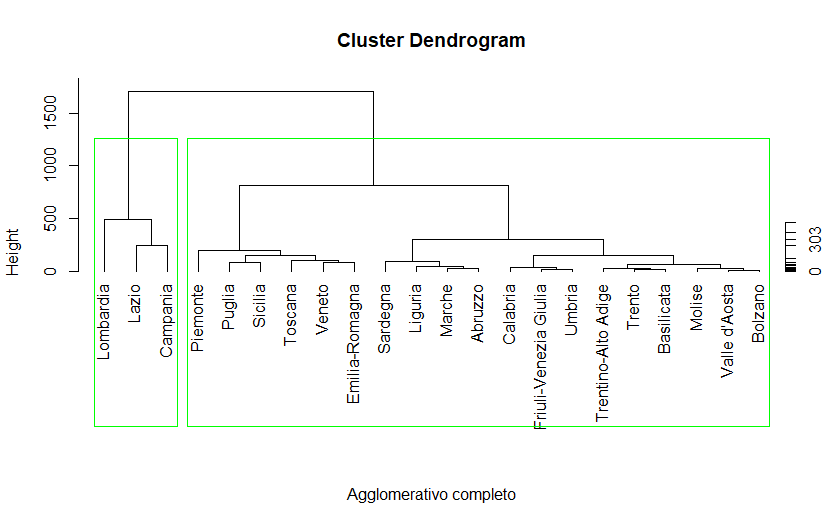
**Metodo del legame completo**

hlc=hclust(d,method = "complete")

plot(hlc,hang=-1,xlab="Agglomerativo completo",sub = " ")

rect.hclust(hlc,k = 2,border = "green")

axis(side=4,at=round(c(0,hls$height),0))



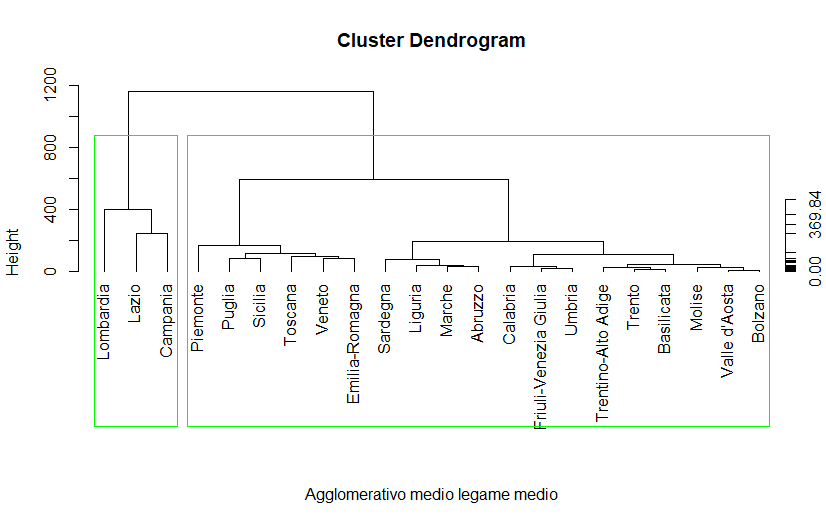
**Metodo del legame medio**

hlm=hclust(d, method="average")

plot(hlm, hang=-1, xlab="Agglomerativo medio legame medio", sub="")

rect.hclust(hlm, k=2, border="green")

axis(side=4, at=round(c(0, hls$height),2))



**Metodo del centroide**

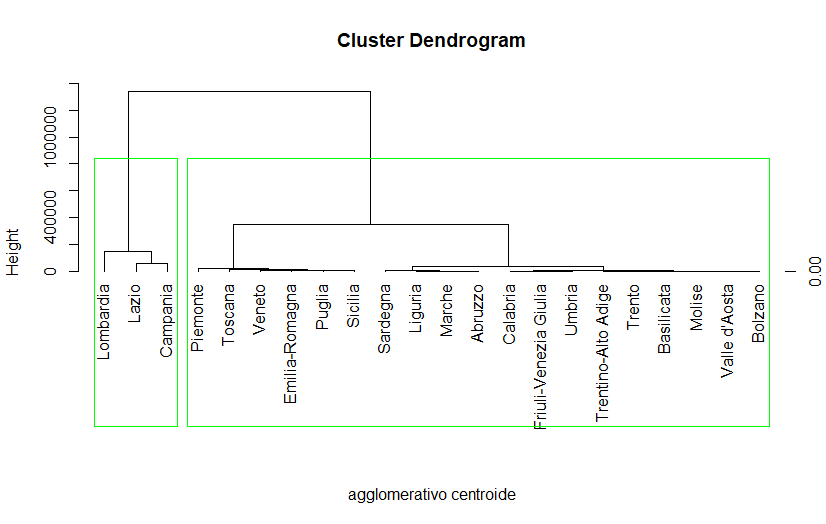
d2=d^2

hc=hclust(d2,method = "centroid")

plot(hc,hang=-1,xlab = "agglomerativo centroide",sub=" ")

rect.hclust(hc,k=2,border = "green")

axis(side=4,at=round(c(0,hls$height),2))



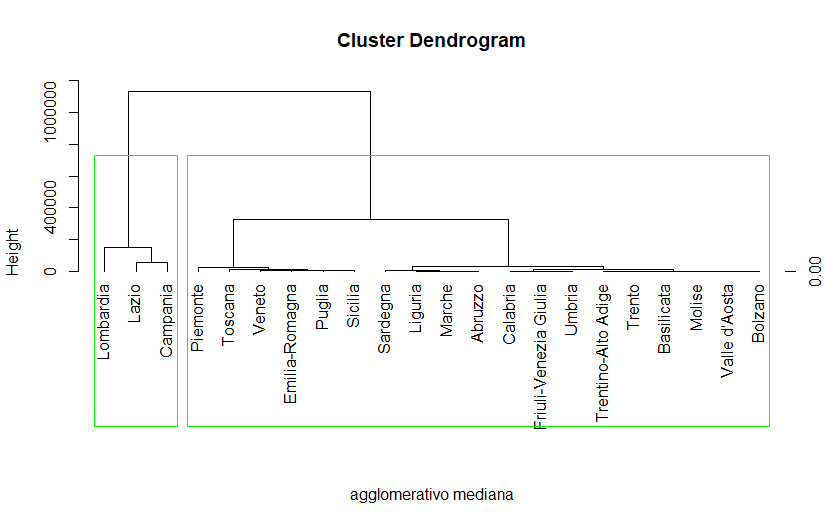
**Metodo della mediana**

hmed=hclust(d2,method = "median")

plot(hmed,hang=-1,xlab = "agglomerativo mediana",sub=" ")

rect.hclust(hmed,k=2,border = "green")

axis(side=4,at=round(c(0,hls$height),2))



Tutti i metodi gerarchici: legame singolo, legame medio, legame completo, metodo del centroide e metodo della mediana hanno fornito il seguente partizionamento in due cluster.

Primo cluster: 19 individui

Secondo cluster: 3 individui

|  |  |
| --- | --- |
| Cluster 2 | Piemonte, Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Veneto, Friuli-Venezia Giulia, Emilia-Romagna, Toscana, Umbria, Marche, Abruzzo, Molise, Puglia, Basilicata, Calabria, Sicilia, Sardegna |
| Cluster 1 | Lombardia, Lazio, Campania |

Si mostra in R il codice per il calcolo delle misure di non omogeneità per i cluster ottenuti con il metodo del legame singolo. Siccome il partizionamento ottenuto è uguale anche per gli altri metodi i risultati saranno uguali.

trh=(n-1)\*sum(apply(z,2,var))

taglio=cutree(hls,k=2)

num=table(taglio)

tagliolist=list(taglio)

agvar=aggregate(z,tagliolist,var)[,-1]

trh1=(num[[1]]-1)\*sum(agvar[1, ])

trh2=(num[[2]]-1)\*sum(agvar[2, ])

trb=trh-trh1-trh2 #misura cluster non omogeeni

rapportolsingolo=trb/trh

La misura di non omogeneità totale trH nel data frame considerato risulta essere uguale a: 5209482

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 1611162.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 131426.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2= 3466893.

Il rapporto =  **0.6654968**.

l metodo non gerarchico K-means ha fornito il seguente partizionamento in due cluster

Primo cluster: 13 individui

Secondo cluster: 9 individui

|  |  |
| --- | --- |
| Cluster 1 | Valle d’Aosta, Liguria, Trentino-Alto Adige, Trento, Bolzano, Friuli-Venezia Giulia, Umbria, Marche, Molise, Basilicata, Calabria, Abruzzo,Sardegna |
| Cluster 2 | Piemonte, Lombardia, Veneto, Emilia-Romagna, Toscana, Lazio, Sicilia, Campania, Puglia |

La misura di non omogeneità totale trT nel data frame considerato risulta essere uguale a 23327101.

La misura di non omogeneità all’interno del primo cluster trH1 risulta essere uguale a 131830.2.

La misura di non omogeneità all’interno del secondo cluster trH2 risulta essere uguale a 1310843.8.

Pertanto, la misura di non omogeneità tra i cluster risulta essere trB=trT-trH1-trH2=16876436.

Il rapporto =  **0.7230677**.

La suddivisione in cluster ottenuta con il metodo non gerarchico K-means risulta essere migliore.

# Variabile aleatoria esponenziale

La **distribuzione esponenziale** è una distribuzione di probabilità continua che descrive la "durata di vita" di un fenomeno che *non invecchia*. Un esempio è la *durata di vita* di una particella radioattiva prima decadere oppure la durata della richiesta di un servizio. Si dice che X ha distribuzione esponenziale di parametro λ>0 e si indica con X∼EXP(λ), se la sua funzione di distribuzione è:

e corrispondente densità di probabilità:

Per una variabile esponenziale si ha che:

Osservando che , se X rappresenta un tempo allora λ rappresenta una frequenza. Quindi se la variabile aleatoria descrive, ad esempio, la durata di vita di un componente elettronico si intuisce che i tempi di vita maggiori corrispondono ai parametri λ più piccoli. Infatti, la funzione densità, al diminuire di λ, si schiaccia sull’asse delle ascisse. Di conseguenza la media si sposta verso valori più elevati e il componente si guasta mediamente più tardi. Pertanto, λ risulta essere inversamente proporzionale al tempo di vita medio del componente.

Come specificato precedentemente, tale variabile aleatoria descrive un fenomeno che non invecchia, ciò significa che è privo di memoria. Gode infatti della seguente proprietà di “assenza di memoria”, per ogni s, t reali positivi risulta:

Se si interpreta X come un tempo di attesa, la precedente equazione mostra che la probabilità condizionata che il tempo di attesa X sia maggiore di t+s dato che essa è maggiore di s non dipende da quanto si è già atteso, ossia da s.

Nel seguente grafico è rappresentata la densità di probabilità e la funzione di distribuzione di una variabile aleatoria con distribuzione esponenziale e parametro λ=3.



La probabilità che la variabile aleatoria esponenziale con λ=3 assuma valori nell’intervallo (0.5, 1.5) è 0.2120212.



La funzione qexp(z, rate) permette di calcolare i quantili:

* Z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile z·100-esimo.
* Rate è il valore del parametro λ.

Il risultato della funzione è il percentile z ·100–esimo, ossia il più piccolo numero x assunto dalla variabile aleatoria esponenziale X tale che

Le seguenti linee di codice permettono di calcolare i quantili di una variabile aleatoria esponenziale con frequenza λ=3.



In R è possibile generare dei campioni casuali utilizzando la funzione rexp. Il seguente codice permette di confrontare la densità teorica esponenziale di parametro λ=3 con la densità simulata generando tre campioni di ampiezza, rispettivamente, 50, 500 e 5000.

png("grafici/densitaEsponenzialeESimulata.png")

par ( mfrow =c(2 ,2))

curve ( dexp(x,rate=3) ,from =0, to=10, xlab="x", ylab="f(x)",main="Densità di probabilità geometrica")

sim<-rexp(50, rate =3)

hist(sim,freq=F,xlim =c(0 ,8) ,ylim =c(0 ,2) ,breaks =100 , xlab ="x", ylab=" Istogramma ",main=" Densita simulata ,N =50 ")

sim<-rexp(500, rate =3)

hist(sim,freq=F,xlim =c(0 ,8) ,ylim =c(0 ,2) ,breaks =100 , xlab ="x", ylab=" Istogramma ",main=" Densita simulata ,N =500 ")

sim<-rexp(5000, rate =3)

hist(sim,freq=F,xlim =c(0 ,8) ,ylim =c(0 ,2) ,breaks =100 , xlab ="x", ylab=" Istogramma ",main=" Densita simulata ,N =5000 ")

dev.off()



Si può notare che all’aumentare dell’ampiezza del campione, l’istogramma delle frequenze relative si avvicina alla densità esponenziale teorica.

## Stima del parametro non noto

Uno dei principali problemi dell’inferenza statistica consiste nello studiare una popolazione descritta da una variabile aleatoria osservabile X di cui si conosce la forma della funzione di distribuzione ma che contiene il valore di un parametro non noto . Per ottenere informazioni sui parametri non noti, si considera un campione di ampiezza n estratto dalla popolazione e si fa uso di alcune variabili aleatorie che sono funzioni misurabili del campione, dette statistiche o stimatori.

Una **statistica** è una funzione misurabile e osservabile del campione casuale . Essendo la statistica osservabile, i valori da essa assunti dipendono soltanto dal campione osservato estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.

Uno **stimatore** è una funzione misurabile e osservabile del campione casuale i cui valori possono essere usati per stimare un parametro non noto della popolazione. I valori assunti da tale stimatore sono detti stime del parametro non noto . Statistiche tipiche sono la media campionaria e la varianza campionaria.

### Stima puntuale

I principali metodi per la stima puntuale sono il metodo dei momenti e il metodo della massima verosimiglianza. Se si hanno k parametri da stimare, il **metodo dei momenti** consiste nell’uguagliare i primi k momenti della popolazione in esame con i corrispondenti momenti del campione casuale. Si tratta quindi di risolvere un sistema di k equazioni in cui i termini a sinistra dipendono dalla legge di probabilità considerata e contengono i parametri non noti, mentre quelli a destra possono essere calcolati a partire dal campione casuale considerato. In particolare, il momento campionario 1-esimo corrisponde alla media campionaria. Il metodo dei momenti fornisce come stimatore del parametro non noto in una variabile esponenziale la media campionaria. Quindi risulta, .

Sia un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza L(𝜗1, 𝜗2, ..., 𝜗k) = L(𝜗1,𝜗2, ..., 𝜗 k; ) del campione osservato () è la funzione di probabilità congiunta (nel caso di popolazione discreta) oppure la funzione densità di probabilità congiunta (nel caso di popolazione assolutamente continua) del campione casuale , ossia:

Il metodo della **massima verosimiglianza** consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri 𝜗1,𝜗2, ...,𝜗k. Anche il metodo della massima verosimiglianza, applicato ad una popolazione esponenziale, fornisce come stimatore del parametro la media campionaria.

Per una popolazione esponenziale la media campionaria è uno stimatore corretto con varianza minima e consistente per 1/ λ.

Uno stimatore del parametro non noto della popolazione è detto **corretto** se e solo se per ogni si ha

ossia se il valore medio dello stimatore è uguale al corrispondente parametro non noto della popolazione.

Uno stimatore si dice corretto e con **varianza uniformemente minima** per il parametro non noto se e solo se per ogni risulta

Uno stimatore del parametro non noto della popolazione è **consistente** se

**Esempio**: Si desidera studiare una popolazione descritta da una variabile aleatoria X con funzione di distribuzione esponenziale. In particolare, il campione ha ampiezza 50 e denota i tempi di gestione in minuti di una richiesta da parte di un servizio A. Si vuole stimare il parametro non noto λ.

Il campione generato con la funzione rexp risulta:

12,55257077 3,557898565 0,066850691 2,412892317

5,634602257 1,380391889 1,946266899 11,42878354

0,480258011 11,3162152 13,66522758 2,618328119

2,881713209 3,709615855 13,45168029 2,040098796

1,843199623 8,041389435 0,675733802 17,35395769

11,3476636 2,758856067 5,484402969 0,625407379

5,265434207 2,992770325 2,284765127 7,652301611

7,803344977 1,518178883 4,278282546 21,26650536

1,026306043 4,415230164 0,762105291 8,245327286

4,940576004 2,101264785 6,98473455 11,59716966

0,070772843 6,792073343 5,675955373 3,50324823

5,427746911 5,56008216 2,798289899 3,364845295

2,041921504 0,87780955

La **stima puntuale** del parametro non noto con il metodo dei momenti e della massima verosimiglianza forniscono come stimatore la media campionaria.

stimatheta <-1.0 /mean (camp)



Risulta quindi λ = 0.1876024.

### Stima intervallare

**Teorema centrale di convergenza**

Sia una successione di variabili aleatorie, definite nello stesso spazio di probabilità, indipendenti ed identicamente distribuite con valore medio µ e varianza σ2 finita e positiva. Posto per ogni intero n positivo , per ogni risulta:

ossia la successione delle variabili aleatorie standardizzate

converge in distribuzione alla variabile aleatoria normale standard.

Il teorema mostra inoltre che sottraendo a la sua media e dividendo la differenza per la deviazione standard di , si ottiene una variabile aleatoria standardizzata la cui funzione di distribuzione è per n sufficientemente grande approssimativamente normale standard. Quindi, per n grande la distribuzione della somma

È approssimativamente normale con valore medio e varianza σ2, ossia

Inoltre, se denotiamo con

la media campionaria, allora

converge in distribuzione alla variabile aleatoria normale standard. Quindi per n grande la distribuzione della media campionaria è approssimativamente normale con valore medio e varianza , ossia

L’approssimazione migliora al crescere di n, nelle applicazioni spesso è già soddisfacente

**Stima intervallare e metodo pivotale**

La **stima intervallare** si propone, a differenza della stima puntuale, di determinare in base ai dati del campione un limite superiore e un limite inferiore entro il quale sia compreso il parametro non noto con un certo coefficiente di confidenza (o grado di fiducia) 1-α.

Un metodo per la costruzione degli intervalli di confidenza è il **metodo pivotale** che consiste nel determinare una variabile aleatoria di pivot che:

* Dipende dal campione casuale ;
* Dipende dal parametro non noto ;
* La sua funzione di distribuzione non contiene il parametro non noto .

Per ogni fissato coefficiente α (0 < α < 1) siano α1 e α2 (α1 < α2) due valori dipendenti soltanto dal coefficiente fissato α e tali che per ogni si abbia:

Se per ogni possibile campione osservato e per ogni si riesce a dimostrare:

con e dipendenti soltanto dal campione osservato allora la relazione

è equivalente a:

Denotando con e con segue che è un intervallo di confidenza di grado per il parametro non noto della popolazione.

**Stima approssimata del parametro non noto di una popolazione esponenziale**

Per effettuare la stima intervallare su un campione con distribuzione esponenziale viene utilizzato il teorema centrale di convergenza. Se X denota la variabile aleatoria che descrive la popolazione con e e con il campione casuale, il teorema centrale di convergenza afferma che la variabile aleatoria

converge in distribuzione ad una variabile aleatoria normale standard. Tale variabile può essere interpretata come una variabile aleatoria di pivot. Pertanto, per campioni di ampiezza elevata è possibile applicare il metodo pivotale in forma approssimata, cioè:

*Sia ) un campione osservato di ampiezza n estratto da una popolazione esponenziale di parametro λ. Se la dimensione del campione è elevata, una stima approssimata dell’intervallo di confidenza di grado 1-α per 1/λ è:*

**Esempio:** Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, determinare una stima dell’intervallo di confidenza di grado 1-α = 0.99 e una stima dell’intervallo di confidenza di grado 1-α = 0.95 per i tempi medi di gestione di una richiesta.

**Stima dell’intervallo di confidenza di grado 1 − α = 0.99**.

alpha <-1 -0.99

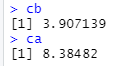
n<-50

m<-5.330421

cb<-m/(1+ qnorm (1- alpha /2,mean =0, sd =1) / sqrt(n))

ca<-m/(1-qnorm (1- alpha /2,mean =0, sd =1) / sqrt(n))

Il limite inferiore risulta cb=3.907139, mentre il limite superiore risulta ca=8.384482.



Risulta quindi:

**Stima dell’intervallo di confidenza di grado 1 − α = 0.95.**

alpha <-1 -0.95

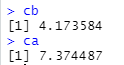
n<-50

m<-5.330421

cb<-m/(1+ qnorm (1- alpha /2,mean =0, sd =1) / sqrt(n))

ca<-m/(1-qnorm (1- alpha /2,mean =0, sd =1) / sqrt(n))

Il limite inferiore risulta cb=4.173584, mentre il limite superiore risulta ca=7.374487.



Risulta quindi:

Si nota che all’aumentare del grado di confidenza 1 – α l’intervallo diventa più grande.

### Confronto tra due popolazioni esponenziali

Consideriamo una prima popolazione esponenziale descritta da una variabile X∼EXP(λ1) con densità di probabilità:

ed una seconda popolazione esponenziale descritta da una variabile Y∼EXP(λ2) con densità di probabilità:

E siano e due campioni casuali di ampiezza n1 e n2 estratti dalle popolazioni esponenziali. Si vuole determinare un intervallo di confidenza di grado 1-α per la differenza 1/ λ 1− 1/ λ 2 per grandi valori di n1 e n2. Dal teorema centrale di convergenza segue che la variabile aleatoria:

Converge in distribuzione ad una variabile aleatoria normale standard, quindi le medie campionarie e sono stimatori corretti e consistenti di 1/ λ 1 e 1/ λ 2, per campioni sufficientemente numerosi l’intervallo di confidenza di grado 1 – α per la differenza 1/ λ 1 e 1/ λ 2 può essere determinato supponendo che:

Da cui:

**Esempio:** Si desidera confrontare i tempi per gestire una richiesta da parte di due servizi A e B. Si supponga che i tempi sono distribuiti come una variabile esponenziale. Il servizio A è descritto da una variabile esponenziale gestisce X∼EXP(λA) 50 richieste, mentre il servizio B è descritto da una variabile esponenziale gestisce Y∼EXP(λB) e gestisce 80 richieste con i seguenti risultati sulle medie e sulle deviazioni standard dei tempi impiegati per ciascuna richiesta: mediaA=5.330421, sdA=4.737098, mediaB=10.11495, sdB=10.82139. Si vuole determinare una stima dell’intervallo di confidenza di grado 1 − α = 0.99 per la differenza 1/ λ A− 1/ λ B tra i tempi medi impiegati per gestire una richiesta da parte dei due servizi.

I valori del campione del servizio A sono quelli elencati precedentemente nel paragrafo 5.1.2. I valori del campione del servizio B sono i seguenti:

12,00437877 30,11981651 36,2732764 5,511409827

3,860684957 33,77324015 40,53770995 6,012060968

1,11152689 0,045442674 3,465351928 0,739132319

36,4652101 53,50107478 4,896664666 3,213116731

7,286642122 4,149451354 11,69270193 3,110677744

15,70766203 23,99411746 1,737911697 6,630810588

10,60050974 3,982419469 1,119041913 6,42280757

14,14385622 3,677816126 2,293871073 7,575114443

2,942939568 9,037694252 27,28819171 0,170129152

5,498995329 7,949043484 9,314060925 1,885408817

4,429491172 2,104222441 20,31249123 28,00607553

5,180053495 3,657424515 7,43016744 14,07186251

1,69316378 1,527658809 3,890972747 4,580235416

4,194758954 9,573003519 3,245679485 13,81225596

1,343266426 1,502164239 6,506694076 1,146984082

13,21266153 13,7555 8,740801522 15,38673783

10,20344139 29,81519982 1,831024233 1,185700023

5,20923879 8,535682289 29,34136988 0,383451576

13,77708594 17,08908137 3,193944894 6,79066407

2,387433276 12,63221709 9,347000354 8,427319665

Il seguente codice permette di calcolare la differenza 1/ λ A− 1/ λ B tra i tempi medi impiegati per gestire una richiesta da parte dei due servizi.

alpha <-1 -0.99

n2<-80

n<-50

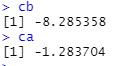
media1<-5.330421

media2<-10.11495

rad<-sqrt(media1^2\*(1/n)+media2^2\*(1/n2))

cb<-media1-media2-qnorm (1-alpha /2, mean=0, sd=1)\*rad

ca<-media1-media2+qnorm (1-alpha /2, mean=0, sd=1)\*rad



Risulta quindi:

Siccome ca e cb sono entrambi negativi, la differenza 1/ λ A − 1/ λ B risulta essere negativa, pertanto λ A > λ B . Siccome in una variabile aleatoria esponenziale λpuò essere visto come una frequenza, il servizio A è in grado di servire più richieste per minuto rispetto al servizio B.

## Verifica delle ipotesi

introduzione su ipotesi nulle ecc.

### Test statistici su grandi campioni

Teoria

Esempi sui 3 tipi di test

**Esempio**: Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, è stato mostrato che una stima dell’intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è . Si si propone di verificare l’ipotesi in alternativa a .

**Esempio**: Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, è stato mostrato che una stima dell’intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è . Si si propone di verificare l’ipotesi in alternativa a .

**Esempio**: Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, è stato mostrato che una stima dell’intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è . Si si propone di verificare l’ipotesi in alternativa a .

### Criterio del chi-quadrato

Teoria

**Esempio**: In 50 osservazioni si riscontra che i tempi medi di gestione della richiesta, da parte di un servizio, espressi in minuti sono di 5.330421. Si desidera verificare utilizzando il test del chi-quadrato se il tempo medio per gestire una richiesta da parte del servizio sia esprimibile con una variabile aleatoria X esponenziale di parametro λ, ossia:

# Bibliografia

|  |  |
| --- | --- |
| [1] | «Che cos'è la violenza di genere,» [Online]. Available: https://www.comune.venezia.it/it/content/cos%C3%A8-la-violenza-di-genere#:~:text=Le%20Nazioni%20Unite%20in%20occasione,minaccia%20di%20tali%20atti%2C%20la. |